

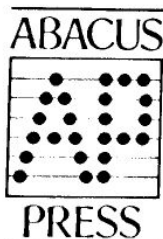
Advanced Physical Geodesy

BY

HELMUT MORITZ



HERBERT WICHMANN VERLAG KARLSRUHE



ABACUS PRESS TUNBRIDGE WELLS KENT

1980

163403

Published simultaneously by Herbert Wichmann Verlag in West Germany
and by Abacus Press in Great Britain, 1980

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Moritz, Helmut:

Advanced physical geodesy / by Helmut Moritz. –
Karlsruhe : Wichmann, 1980. –
(Sammlung Wichmann : N.F. : Buchreihe ; Bd. 13)
ISBN 3-87907-106-3

ISBN 3-87907-106-3

© 1980 Herbert Wichmann Verlag, Rheinstraße 122, 7500 Karlsruhe 21,
West Germany

ISBN 0 85626 195 5

Abacus Press, Abacus House, Speldhurst Road, Tunbridge Wells,
Kent TN4 0HU, England

All rights reserved. No part of this publication may be reproduced, stored
in a retrieval system, or transmitted in any form or by any means,
electronic, mechanical or otherwise, without prior permission of
Herbert Wichmann Verlag.

Printed and bound in the Federal Republic of Germany.

To the memory of

RONALD SUNTHERERAJ MATHER

1933 - 1978

PREFACE

Physical geodesy, the study of the gravitational field and the figure of the earth, has seen enormous progress in the years following the publication of the book "Physical Geodesy" by W.A. Heiskanen and the present author in 1967. The new book is devoted almost exclusively to this progress, but even so it is far from comprehensive. First, it is limited to the mathematical theory, of which it attempts a systematic and didactic presentation. There is hardly any mention of observational techniques and of numerical results, but the mathematical methods are developed with a view to practical application.

Secondly, even from the theory of physical geodesy, a selection had to be made to keep the size of the book within limits. Least-squares collocation, which is a technique for combining observational data of different types for an optimal determination of the earth's figure and gravitational field, is treated rather broadly. The elementary presentation in Part B should be sufficient for most practical applications, whereas Part C provides the advanced theory which is necessary for a deeper understanding. Part D deals with the geodetic boundary-value problem, the problem of Molodensky, but limited to two main topics: series solutions as proposed by Molodensky, Brovar, and others, which seem to be most convenient for practical use; and recent mathematical investigations regarding existence and uniqueness of the solution, associated with the names of Hörmander, Krarup, and Sansø.

The book is restricted to what might be called "classical physical geodesy": both the figure of the earth and its gravitational field are considered independent of time. This is true to a very high accuracy (almost down to 10^{-7}), which is sufficient for most present applications. For higher accuracy, geodynamical (time-dependent) effects can be taken into account by small corrections.

This approach seems to be practically and didactically the best; it has so far almost exclusively been pursued. Thus the present book uses it too, rather than formulating the observation equations and the boundary-value problem from the very beginning in a temporally variable, "four-dimensional", form.

An adequate treatment of geodynamical effects would have required a special Part E, if not another book. The handling of this topic within the frame of a section (sec.55) is a meager substitute: only the barest outlines could be sketched.

The author also regrets not to have been able to include a treatment of the differential structure of the gravity field, again for reasons of space. There is, however, the book by Hotine (1969), which serves as an excellent basis for a study of the extensive subsequent literature.

The understanding of the book requires a basic knowledge of physical geodesy. For the sake of uniformity, we have used the text (Heiskanen and Moritz, 1967) as a source of reference. However, an equally useful background is provided by books such as (Groten, 1979), (Ledersteger, 1969), (Levallois, 1970), (Magnizki et al., 1964), (Pellinen, 1978), (Pick et al., 1973), (Shimbirev, 1975), or (Torge, 1975).

The book is written for graduate students and research workers in the field of geodesy and gravity; it is not a mathematical text. In fact, intuitive intelligibility is aimed at, rather than full abstract rigor. The author has been guided by a kind of "minimum principle": to present the topics with the minimum adequate mathematical apparatus. Still, some places do require quite advanced mathematics, of which an easygoing introduction is given in Part A. Throughout, we have provided ample, wordy, and sometimes repetitive explanations, because the mathematical and physical meaning behind the formulas is as important as the formulas themselves. Also, derivations are usually presented in a quite detailed manner.

We have tried to use a fairly uniform notation without being pedantic. Vectors and matrices have been symbolized by underlined letters where necessary to avoid confusion; otherwise ordinary letters are employed for denoting them. Similarly, row and column vectors are distinguished only where matrix operations are involved.

The list of references is intended as a guide for the reader rather than as a complete documentation, which would comprise much more than the 200 titles given. Without doubt there are important omissions due to the author's inadvertence or ignorance. He apologizes to any colleague who feels that his work has not been adequately represented.

The author's cooperation, now for almost two decades, with the Department of Geodetic Science of The Ohio State University has been of invaluable influence on his research. He is particularly indebted to discussions with Dr. Richard H. Rapp, Mr. Béla Szabó, and Dr. Urho A. Uotila, who has also given permission for the frequent use of material from reports of this department.

The Editors of the Bollettino di Geodesia e Scienze Affini and of the publications of the German Geodetic Commission also have kindly permitted the use of some material published there.

The author expresses particular thanks to Mrs. Astrid Fink-Gradl for the competent and painstaking preparation of the typescript in a form suitable for direct reproduction and for advice in linguistic questions, to Mr. Robert Geretschläger for properly constructing the diagrams, and to him and to Dr. Hans Sünkel for help in proofreading.

Graz, Austria, November 1979

Helmut Moritz

TABLE OF CONTENTS

PART A

GENERAL BACKGROUND

1. The Earth's Gravity Field	2
2. Reference Ellipsoid and Anomalous Gravity Field	7
3. Spherical Harmonics	18
4. A First Look at Hilbert Space	24
5. Normed Spaces	40
6. Convergence of Spherical Harmonics I	50
7. Convergence of Spherical Harmonics II	63
8. Runge's Theorem	67

PART B

LEAST-SQUARES COLLOCATION: ELEMENTARY APPROACH

9. Least-Squares Prediction	76
10. The Covariance Function	81
11. Least-Squares Collocation	84
12. Invariance Properties; Analytical Collocation	91
13. Application to Bjerhammar's problem	95
14. Collocation with Random Errors	99
15. Application to Geoid Determination	106
16. Least-Squares Collocation with Parameters	111
17. Accuracy	122
18. Application to Physical Geodesy	132
19. Stepwise Collocation	144
20. Accuracy in Stepwise Collocation	150
21. Determination of Spherical Harmonics	156
22. Local Structure of Covariance Functions	169
23. Global Covariance Models	181

PART C

LEAST-SQUARES COLLOCATION: ADVANCED ASPECTS

24. Hilbert Spaces with Kernel Functions	196
25. Collocation and Hilbert Space	207
26. Geodetic Measurements and Their Representation	221
27. Linearization	230
28. Variational Principles	238
29. Solution of a Variational Problem	243
30. Least-Squares Collocation and Related Models	249
31. Stochastic Processes on the Circle	260
32. The Covariance Function	263
33. Ergodic Processes on the Circle	269
34. Stochastic Processes on the Sphere	279
35. Ergodic Processes on the Sphere	285
36. Rotation Group Space	288
37. Statistical Distributions in Rotation Group Space	297
38. The Meaning of Statistics in Collocation	307
39. Ellipsoidal Corrections	314

PART D

THE GEODETIC BOUNDARY-VALUE PROBLEM

40. Molodensky's Problem	330
41. Linearization	336
42. Spherical Approximation	349
43. Molodensky's Solution	354
44. Brovar's Solution	365
45. Solution by Analytical Continuation	377
46. Pellinen's Equivalence Proof	388
47. Convergence of Molodensky's Series	401
48. Use of the Terrain Correction	414
49. Practical Aspects	419
50. Existence and Uniqueness for the Linearized Molodensky Problem	428
51. Hörmander's Results for the Nonlinear Problem	434
52. The Gravity Space Approach	449

53. Linearization	457
54. Sansô's Treatment of the Nonlinear Problem	467
55. Geodynamical Effects	477
References	490
Index	498

PART A

GENERAL BACKGROUND

This introductory part is intended mainly to provide the geodetic and mathematical background for the present book. Sections 1 to 3 review essential material from the theory of the earth's gravity field, including spherical harmonics. Basic concepts from functional analysis, such as linear operators and functionals, will be used throughout the book. Therefore, sections 4 and 5 give a simple introduction to these topics, which is intended for geodesists, not for mathematicians.

The following sections are more advanced. Section 6 presents a review of the difficult problem of convergence of the spherical harmonic expansion of the external gravitational potential at the earth's surface. A fresh look on the subject from a practical angle is provided by an application of Runge's theorem in sec. 7. This theorem is proved in sec. 8, which is mathematically more demanding than the preceding sections.

The reader may start with sections 1 to 3 to refresh his knowledge of basic facts from physical geodesy and get familiar with the terminology. If he is interested in applications rather than in the theory, he may then pass on to Part B.

For the study of Parts C to E, the material of sections 4 and 5 is indispensable. Runge's theorem (sec.8) can be studied when the need arises; the proof (p.70 et seq.) may be left out. Sections 6 and 7 may be read whenever desired.

2 General Background

1. THE EARTH'S GRAVITY FIELD

This section reviews basic properties of the earth's gravity field and coordinate systems related to it, in general following (Heiskanen and Moritz, 1967), especially sections 1-1 and 2-1 through 2-4.

Our fundamental earth-fixed rectangular coordinate system xyz is defined in the usual way: the origin is at the earth's center of mass (the *geocenter*); the z -axis coincides with the mean axis of rotation, the x -axis lies in the mean Greenwich meridian plane and is normal to the z -axis; the y -axis is normal to the xz -plane and directed so that the xyz system is right-handed; the xy -plane is thus the (mean) equatorial plane.

One uses a mean axis of rotation and a mean Greenwich meridian plane in order to get a definition independent of time, in view of very small and more or less periodic changes in the instantaneous rotation axis and of deformations of the earth's body; see sec. 55.

The *gravitational potential* V may be expressed by the formula

$$V(P) = V(x,y,z) = G \iiint_{\text{earth}} \frac{\rho(Q)}{r} dv_Q, \quad (1-1)$$

where P is a point having coordinates (x,y,z) , Q is a point, variable within the earth's body, which forms the center of the volume element dv_Q ,

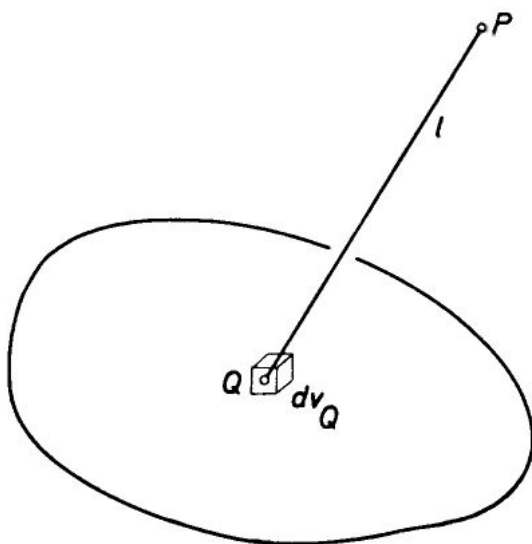


FIGURE 1.1. Illustrating equation (1-1).

r is the distance between P and Q , and $\rho(Q)$ is the mass density at Q ;
 G is the Newtonian gravitational constant

$$G = 6.672 \times 10^{-11} \text{ m}^3 \text{ s}^{-2} \text{ kg}^{-1} . \quad (1-2)$$

The integral is to be extended over the whole earth's body, which includes the solid and liquid parts. The (very small) effect of the atmosphere is usually disregarded; if necessary, it can be taken into account by corrections, which have the relative order of 10^{-6} . The same treatment may be applied to temporal variations of V , which have the order of 10^{-7} ; see secs. 49 and 55. Unless stated otherwise, we shall therefore treat the earth as a rigid body without temporal changes and without atmosphere.

Even so, the representation (1-1) has only theoretical value because its practical use would require the knowledge of the detailed density distribution within the earth, which obviously is not known.

For large distances

$$r = \sqrt{x^2 + y^2 + z^2} ,$$

(1-1) may be expressed as

$$V = \frac{GM}{r} + O\left(\frac{1}{r^2}\right) \quad \text{as } r \rightarrow \infty . \quad (1-3)$$

M denoting the total mass of the body and $O(1/r^2)$ symbolizing a term that, for $r \rightarrow \infty$, tends to zero as $1/r^2$. The physical sense of this equation is that, at large distances and approximately, any body acts gravitationally as a point mass.

The *gravity potential* W is the sum of V and the potential of the centrifugal force,

$$V_c = \frac{1}{2} \omega^2 (x^2 + y^2) , \quad (1-4)$$

so that

$$W(x, y, z) = V(x, y, z) + \frac{1}{2} \omega^2 (x^2 + y^2) , \quad (1-5)$$

ω being the angular velocity of the earth's rotation (which is considered constant).

The field of potential V is called the *gravitational field*; the field of potential W is the *gravity field*.

4 General Background

The *gravity vector* \underline{g} is the gradient of W :

$$\underline{g} = \text{grad } W = \begin{bmatrix} W_x \\ W_y \\ W_z \end{bmatrix} ; \quad (1-6)$$

its components are the partial derivatives of W with respect to x, y, z ; it is the resultant of the gravitational force $\text{grad } V$ and the centrifugal force.

The second-order partial derivatives of V form a symmetric matrix

$$\begin{bmatrix} V_{xx} & V_{xy} & V_{xz} \\ V_{yx} & V_{yy} & V_{yz} \\ V_{zx} & V_{zy} & V_{zz} \end{bmatrix} , \quad (1-7)$$

which is called the (second-order) *gravitational gradient tensor*. Similarly, the second-order derivatives of W form the *gravity gradient tensor*.

The trace of the matrix (1-7) is the *Laplacian* of V :

$$\Delta V = V_{xx} + V_{yy} + V_{zz} . \quad (1-8)$$

Outside the attracting masses, above the earth's surface S , V satisfies *Laplace's equation*

$$\Delta V = 0 ; \quad (1-9)$$

the solutions of this equation are called *harmonic functions*. In the earth's interior, inside S , the potential V satisfies *Poisson's equation*

$$\Delta V = -4\pi G\rho , \quad (1-10)$$

ΔV and ρ referring to the same point inside S .

The corresponding relations for the gravity potential W are, in view of (1-4),

$$\Delta W = 2\omega^2 \quad \text{outside } S , \quad (1-11)$$

$$\Delta W = -4\pi G\rho + 2\omega^2 \quad \text{inside } S . \quad (1-12)$$

The magnitude, or norm, of the gravity vector \underline{g} is gravity g :

$$g = ||\underline{g}|| ; \quad (1-13)$$

the direction of \underline{g} , expressed by the unit vector

$$\underline{n} = -g^{-1}\underline{g} , \quad (1-14)$$

is the direction of the vertical, or plumb line; we have chosen the minus sign so that \underline{n} points upwards.

In our basic xyz system, the vector \underline{n} has components which can be expressed in terms of two angles ϕ , Λ as

$$\underline{n} = \begin{bmatrix} \cos\phi \cos\Lambda \\ \cos\phi \sin\Lambda \\ \sin\phi \end{bmatrix} . \quad (1-15)$$

The angles defined in this way are called the *astronomical coordinates*: astronomical latitude ϕ and astronomical longitude Λ .

Fig. 1.2 illustrates these astronomical coordinates by means

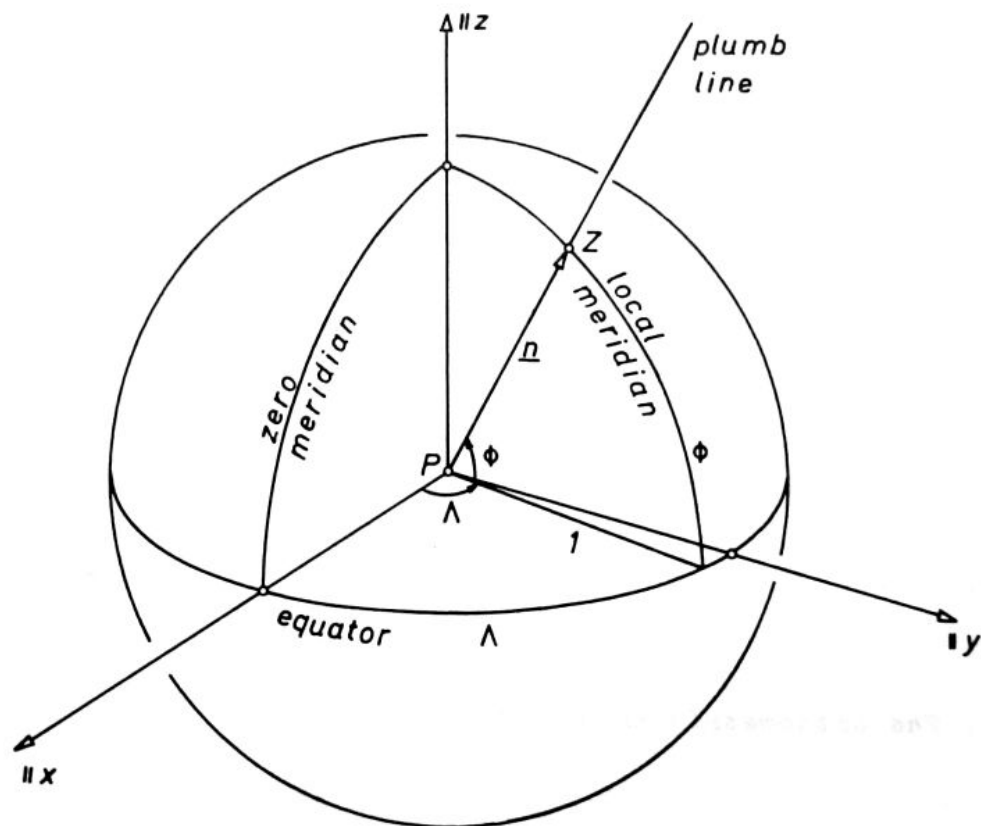


FIGURE 1.2. Astronomical latitude ϕ and longitude Λ on the unit sphere.

of a unit sphere centered at the observation station P . The symbols $||x$, $||y$, $||z$ denote parallels, through P , to the coordinate axes. The intersection of the sphere by the planes $(||x, ||y)$ and $(||x, ||z)$ is the equator and the zero meridian, respectively. The local plumb line intersects the sphere at the zenith Z ; the vector PZ is the unit vector \underline{n} . The coordinates ϕ and λ appear as angles, as well as arcs on the unit sphere.

The surfaces $W = \text{const.}$ are called the *equipotential surfaces* or *level surfaces*. They are everywhere normal to the gravity vector, that is, to the plumb line. A particular one of these surfaces,

$$W(x,y,z) = W_0 = \text{const.}, \quad (1-16)$$

which approximately forms an average surface of the oceans, is distinguished by calling it the *geoid*.

The orthogonal trajectories of the level surfaces are the lines of force. The tangent to a line of force at any of its points is the direction of the gravity vector \underline{g} , or the plumb line. Sometimes the lines of force, which are slightly curved, are themselves referred to as plumb lines; a confusion is not likely to arise.

Let P be a point of the visible earth's surface, called the *topographic surface* or the *physical earth's surface*. The line of force passing through P intersects the geoid at a point P_0 . The length of the (slightly curved) plumb line segment P_0P is the *orthometric height* H .

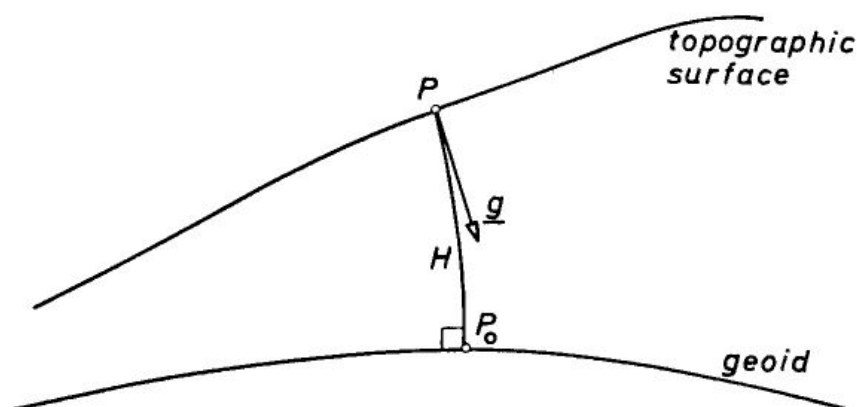


FIGURE 1.3. The orthometric height H .

The triple (ϕ, λ, H) is called the *natural coordinates* of P . They form a system of curvilinear coordinates defined in terms of the gravity field.

An alternative definition of natural coordinates is by the triple (ϕ, λ, W) , since the potential W of P can also be regarded as a physical measure of the elevation of P . This is particularly evident if we consider the *geopotential number*

$$C = W_0 - W, \quad (1-17)$$

which is easily related to H but is conceptually simpler.

Finally it should be noted that physical geodesy is concerned almost exclusively with the gravity field at the geoid and above. Of particular interest is the external gravitational field, the field outside the earth's surface, for which the potential V is a harmonic function.

2. REFERENCE ELLIPSOID AND ANOMALOUS GRAVITY FIELD

This is again a review section, summarizing some basic material which is presented in detail, for instance, in chapters 2 and 5 of (Heiskanen and Moritz, 1967).

Geodetic coordinates. If we replace the geoid by an ellipsoid, then the natural coordinates ϕ, λ, H are replaced by the geodetic coordinates ϕ, λ, h . They are defined in the following way (Fig. 2.1).

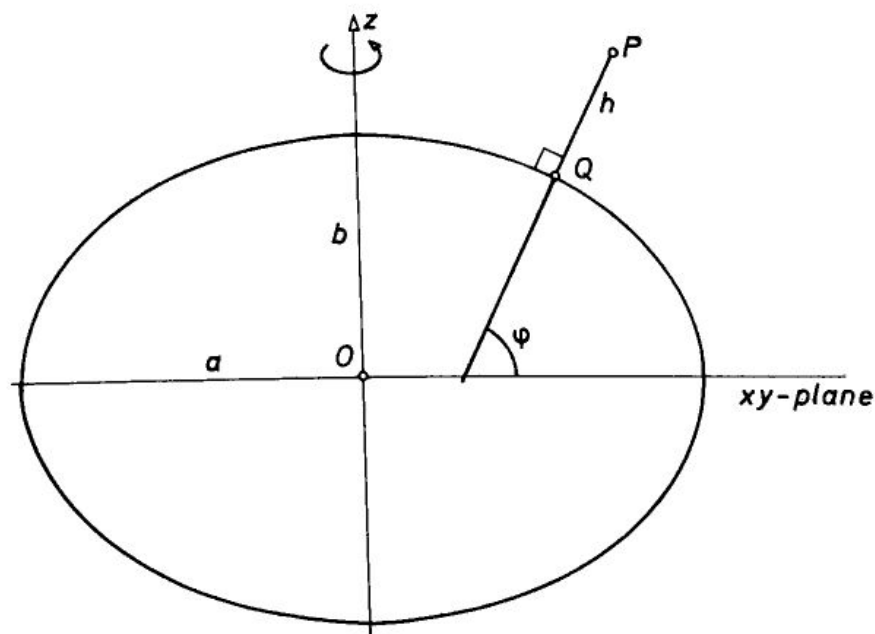


FIGURE 2.1. Reference ellipsoid and geodetic coordinates.

An ellipsoid of revolution, generated by rotating an ellipsoid of semi-axes a and b about the minor axis, is placed with its center at the geocenter, in such a way that the minor axis coincides with the z -axis. A spatial point P is projected, by means of the straight line normal to the ellipsoid, onto the ellipsoid; this gives the point Q . The straight segment QP is the *geodetic height* h , and the usual ellipsoidal geographical coordinates of the foot point Q are the *geodetic latitude* ϕ and *geodetic longitude* λ of P . More precisely, ϕ is the angle between the ellipsoidal normal and the equatorial plane, which is the xy -plane, and λ is the angle between the meridian plane of P (the plane through P and the z -axis) and the zero meridian plane, which is the xz -plane.

The system (ϕ, λ, h) is related to the system (x, y, z) by closed formulas:

$$\begin{aligned} x &= (v + h) \cos \phi \cos \lambda, \\ y &= (v + h) \cos \phi \sin \lambda, \\ z &= \left(\frac{b^2}{a^2} v + h\right) \sin \phi, \end{aligned} \quad (2-1)$$

where

$$v = \frac{c}{\sqrt{1 + e'^2 \cos^2 \phi}}, \quad (2-2)$$

$$c = \frac{a^2}{b}, \quad (2-3)$$

$$e'^2 = \frac{a^2 - b^2}{b^2}; \quad (2-4)$$

v is the transversal radius of curvature of the ellipsoid, c is the polar radius of curvature, and e' is called the second (numerical) excentricity.

The deviation between the plumb line and the ellipsoidal normal is characterized by two small angles ξ, n , the components of the *deflection of the vertical*. This is illustrated by Fig. 2.2. The (astronomical) zenith Z is the spherical image of the plumb line and has the spherical coordinates ψ and Λ , as in Fig. 1.2. The "geodetical zenith" Z' is the image of the ellipsoidal normal and has the coordinates ϕ and λ . From Fig. 2.2 it follows immediately that

$$\begin{aligned} \xi &= \psi - \phi, \\ n &= (\Lambda - \lambda) \cos \phi. \end{aligned} \quad (2-5)$$

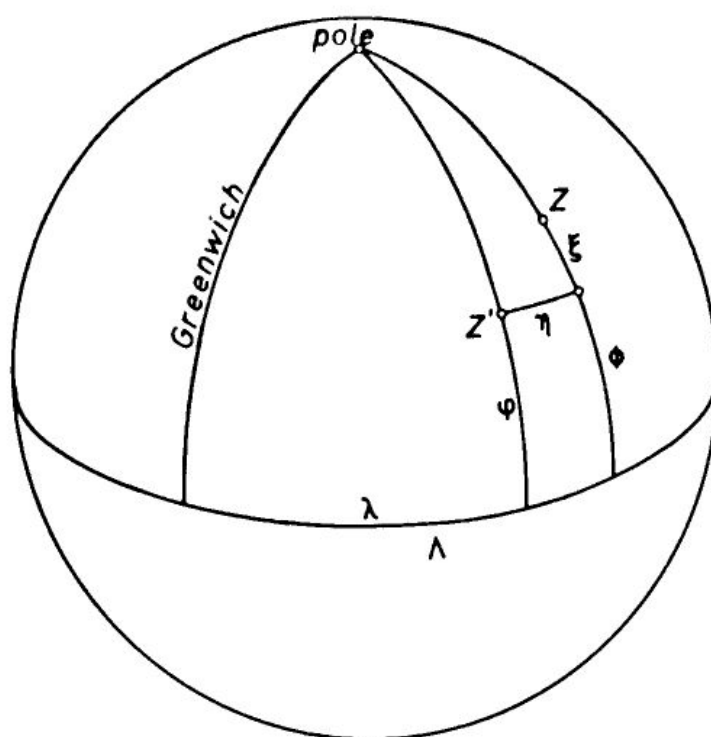


FIGURE 2.2. The deflection of the vertical on the unit sphere.

Similarly, Fig. 2.3 shows that to a sufficient approximation, N being the geoidal height, we have

$$N = h - H . \quad (2-6)$$

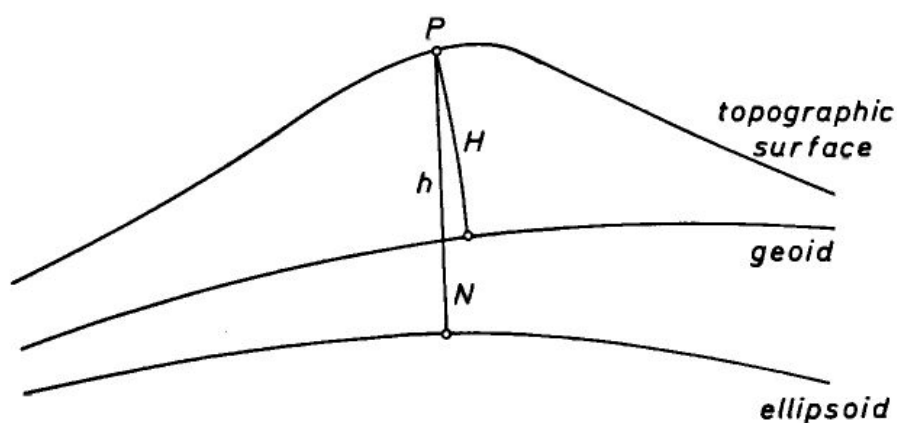


FIGURE 2.3. The geoidal height.

Equations (2-5) and (2-6) relate the geodetic and the natural coordinates through the deflection of the vertical and the geoidal height.

The normal gravity field. The ellipsoid may be considered as some "normal surface" for the geoid: a suitably chosen ellipsoid of revolution closely approximates the geoid and represents its global shape (the maximum deviation of the geoid from such a best-fitting ellipsoid is on the order of only 100 m !). It is, therefore, natural to use the external gravity potential of an ellipsoid of revolution as a *normal gravity potential* U to approximate the earth's external gravity potential W .

Since the geoid

$$W(x,y,z) = W_0 = \text{const.} \quad (2-7)$$

is an equipotential surface of W , it is obvious to postulate that the ellipsoid be an equipotential surface for U :

$$U(x,y,z) = U_0 = \text{const.}, \quad (2-8)$$

so that the given ellipsoid becomes an *equipotential ellipsoid*, or *level ellipsoid*. Furthermore, U must be the sum

$$U(x,y,z) = V(x,y,z) + \frac{1}{2} \omega^2 (x^2 + y^2) \quad (2-9)$$

of a normal gravitational potential V and a centrifugal potential; V must satisfy Laplace's equation

$$\Delta V = 0 \quad (2-10)$$

outside the ellipsoid and behave at infinity approximately as a point mass:

$$V = \frac{GM}{r} + O\left(\frac{1}{r^2}\right) \quad \text{as } r \rightarrow \infty, \quad (2-11)$$

M denoting the total mass enclosed by the ellipsoid. These equations correspond to (1-5), (1-9), and (1-3), respectively.

It can be shown that the postulate (2-8), together with the natural conditions (2-9), (2-10), and (2-11), completely and unambiguously determine the normal gravity potential U , provided the numerical values of

$$\begin{aligned} a, b \dots & \text{ semiaxes of the ellipsoid,} \\ \omega \dots\dots & \text{ angular velocity of rotation,} \\ U_0 \dots\dots & \text{ normal potential at the ellipsoid,} \end{aligned} \quad (2-12)$$

or of four other suitable constants, are given.

The function U defined in this way is expressed by a closed formula which, however, involves some new notations and will not be used in this book. We shall, therefore, not give this formula here and limit ourselves to stating some auxiliary relations, referring the reader for details, e.g., to (Heiskanen and Moritz, 1967, secs. 2-7 to 2-9).

The mass of the ellipsoid is expressed by

$$GM = \frac{E}{\arctan e'} (U_0 - \frac{1}{3} \omega^2 a^2) , \quad (2-13)$$

where

$$E = \sqrt{a^2 - b^2} \quad (2-14)$$

is the linear excentricity and

$$e' = \frac{E}{b} \quad (2-15)$$

is the second (numerical) excentricity of the ellipsoid, which we already have met at (2-4).

Normal gravity γ on the ellipsoidal surface is given by the formula of Somigliana

$$\gamma = \frac{a\gamma_a \cos^2 \phi + b\gamma_b \sin^2 \phi}{\sqrt{a^2 \cos^2 \phi + b^2 \sin^2 \phi}} , \quad (2-16)$$

where ϕ is the geodetic latitude and where normal gravity at the equator, γ_a , and at the poles, γ_b , are given by

$$\gamma_a = \frac{GM}{ab} \left(1 - m - \frac{me'q'_0}{6q_0} \right) , \quad (2-17)$$

$$\gamma_b = \frac{GM}{a^2} \left(1 + \frac{me'q'_0}{3q_0} \right) ,$$

with

$$m = \frac{\omega^2 a^2 b}{GM} , \quad (2-18)$$

$$q_0 = \frac{1}{2} \left(1 + \frac{3}{e'^2} \right) \arctan e' - \frac{3}{2e'} , \quad (2-19)$$

$$q_0' = 3\left(1 + \frac{1}{e'^2}\right)\left(1 - \frac{1}{e'} \arctan e'\right) - 1. \quad (2-20)$$

For later use we shall need the quantity

$$J_2 = \frac{C - A}{Ma^2}, \quad (2-21)$$

where A and C are the principal moments of inertia of our ellipsoid of revolution: A is the moment about the x (or y) axis, and C is the moment about the z -axis. This quantity J_2 , called the *dynamic form factor*, is, for the level ellipsoid, expressed by

$$J_2 = \frac{1}{3} e^2 \left(1 - \frac{2me'}{15q_0}\right) \quad (2-22)$$

(*loc.cit.*, p.73), where

$$e = \frac{E}{a} \quad (2-23)$$

is the first (numerical) excentricity.

The reader should note that all these quantities are, in fact, expressed in terms of the four constants (2-12).

The anomalous gravity field. The normal gravity potential U is a good first approximation for the actual gravity potential W outside the geoid. The difference

$$T = W - U, \quad (2-24)$$

called the *anomalous potential*, or *disturbing potential*, is small. The function T is harmonic outside the earth, satisfying Laplace's equation

$$\Delta T = 0; \quad (2-25)$$

this follows by forming the difference between (1-11) and the corresponding equation for U , under the assumption that ω is the same in both cases.

Writing

$$W = U + T, \quad (2-26)$$

we see that the gravity field (potential W) is split up into a normal field (potential U) and an anomalous field (potential T). This decomposition is very practical: the principal part, expressed by U , is given by closed (ellipsoidal) formulas; the remainder, expressed by T , is irregular but very small, so that linear approximations are sufficient in practice.

These principles, decomposition and linear formulas, will now be demonstrated for quantities referring to geoid and reference ellipsoid, respectively (Fig. 2.4).

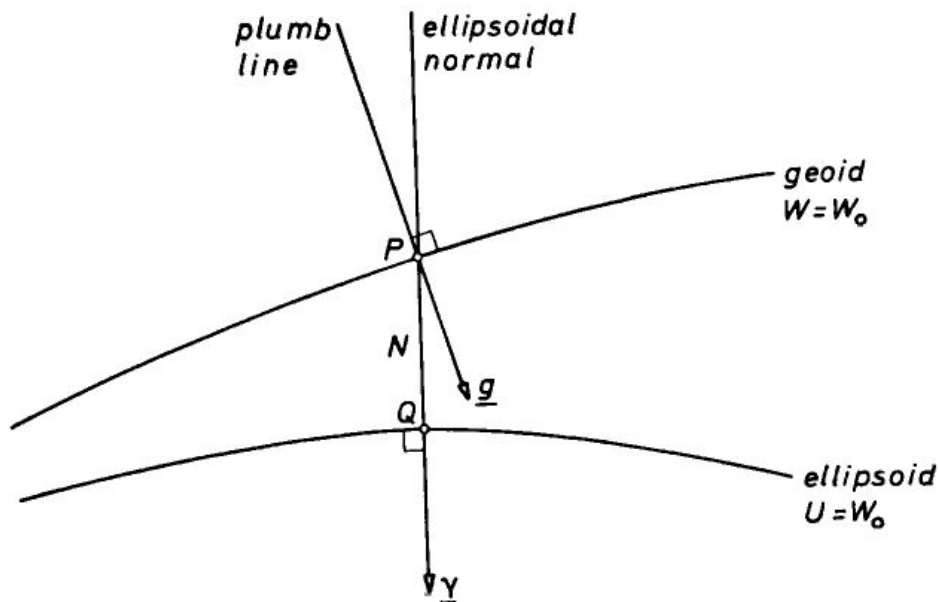


FIGURE 2.4. Geoid and reference ellipsoid.

Let us associate to each geoidal point P an ellipsoidal point Q by projecting P onto the ellipsoid by means of the ellipsoidal normal. The distance PQ is the geoidal height N already introduced above. The normal gravity vector $\underline{\gamma}$ at Q ,

$$\underline{\gamma} = (\text{grad } U)_Q, \quad (2-27)$$

is considered to be the "normal" counterpart of the gravity vector \underline{g} at P ,

$$\underline{g} = (\text{grad } W)_P . \quad (2-28)$$

The difference between the norms g and γ of these vectors is the *gravity anomaly*

$$\Delta g = g - \gamma ; \quad (2-29)$$

note that gravity g refers to P whereas normal gravity refers to Q . On the other hand, in (2-24) both W and U refer to the same point: both to P or both to Q .

The equations (2-5), (2-6), (2-24), and (2-29) all have the same structure: quantities of the anomalous gravity field ($T, N, \Delta g, \xi, \eta$) are expressed in terms of differences between actual field quantities (W, H, g, ϕ, λ) and their normal or ellipsoidal counterparts ($U, h, \gamma, \phi, \lambda$).

Basic is the fact, already mentioned, that all relations between quantities of the anomalous gravity field are *linear*, obtained by Taylor expansions truncated after the linear term. We mention the most important relationships. They have a particularly simple form if the normal potential U_0 at the ellipsoid is taken to be equal to the actual potential W_0 at the geoid; this will be assumed.

The components ξ, η of the deflection of the vertical are connected with the geoidal height N by

$$\xi = - \frac{\partial N}{\partial u} , \quad \eta = - \frac{\partial N}{\partial v} \quad (2-30)$$

where the system uv is a local cartesian coordinate system in the tangent plane to the geoid at P , with origin at P , the u -axis pointing north and the v -axis pointing east (cf. Heiskanen and Moritz, 1967, p.112). The geoidal height N is related to the anomalous potential T by *Bruns' formula*

$$N = \frac{T}{\gamma} , \quad (2-31)$$

where $T = W_Q - U_Q$ (or, to the same accuracy, $= W_P - U_P$), and T and Δg are related by

$$\frac{\partial T}{\partial h} - \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} T + \Delta g = 0 , \quad (2-32)$$

where $\partial/\partial h$ denotes differentiation along the ellipsoidal normal (*ibid.*, pp.85-86).

The last formula is called the *fundamental equation of physical geodesy*. It is a boundary condition at the ellipsoid. By solving Laplace's equation (2-25) subject to the boundary condition (2-32), supposing Δg to be given, one obtains T outside and at the ellipsoid. Then Bruns' formula (2-31) yields the geoidal height.

The validity of $\Delta T = 0$ outside the ellipsoid presupposes the unrealistic situation that there are no masses outside the ellipsoid. Therefore, these masses must be removed computationally, by a so-called *gravity reduction*. (A logically more satisfactory approach, due to Molodensky, will be considered in Part D.)

An explicit solution for T in terms of the boundary values Δg is found in the following way. Since the quantities T and Δg entering in (2-32) are very small, the flattening f may be neglected in this equation (this introduces an error in T of $fT \approx 0.003 T$). Then this boundary condition takes the form

$$\frac{\partial T}{\partial r} + \frac{2}{R}T + \Delta g = 0 \quad , \quad (2-33)$$

where

$$R = 6371 \text{ km} \quad (2-34)$$

is a mean radius of the earth. Thus the reference ellipsoid is formally replaced by a sphere of radius R , and $\partial/\partial r$ is the radial derivative at this sphere.

The meaning of this *spherical approximation* should be properly understood. It does not mean that the reference ellipsoid is replaced by a sphere in a geometrical sense, so that now a sphere, instead of an ellipsoid, would be used as a reference surface for the geoid. It only means that the flattening is neglected in the coefficients of ellipsoidal formulas such as (2-32), so that *formally* a spherical relation (2-33) is obtained. As a matter of fact, normal gravity γ in $\Delta g = g - \gamma$ must be computed by the exact ellipsoidal formula (2-16).

The solution of Laplace's equation $\Delta T = 0$ subject to the boundary condition (2-33) leads to *Stokes formula*

$$T_P = \frac{R}{4\pi} \iint_{\sigma} \Delta g_Q S(\psi) d\sigma \quad , \quad (2-35)$$

where P is the point at which T is computed and Q is the variable point to which Δg refers. The notations are illustrated by means of

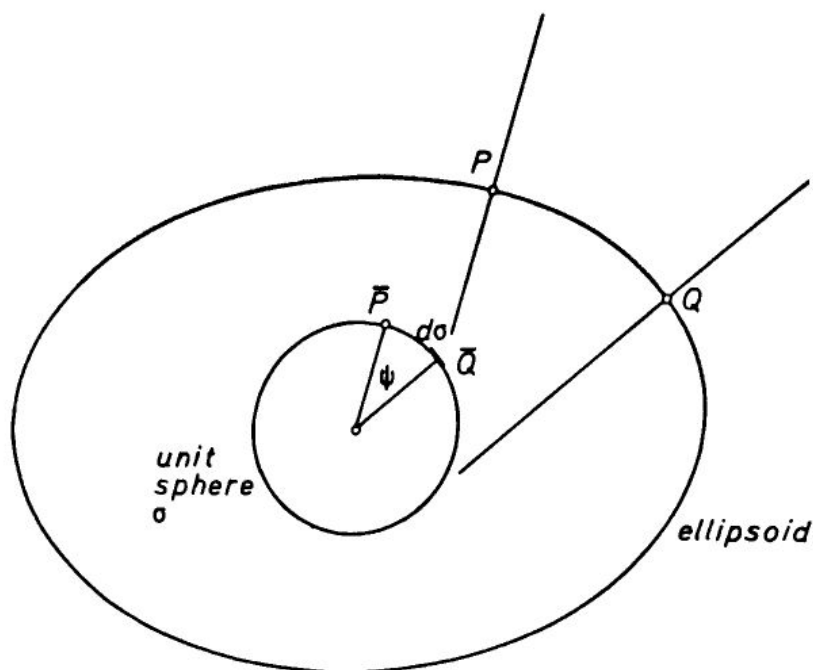
FIGURE 2.5. *Reference ellipsoid and unit sphere.*

Fig. 2.5. The symbol σ denotes a unit sphere whose center may be taken to coincide with the center of the ellipsoid. The points P and Q are mapped into the points \bar{P} and \bar{Q} on the sphere σ in such a way that the radii at \bar{P} and \bar{Q} are parallel to the ellipsoidal normals at P and Q ; that is, the spherical coordinates (latitude and longitude) of \bar{P} and \bar{Q} are identified with the geodetic coordinates ϕ, λ of P and ϕ', λ' of Q , respectively. The surface element $d\sigma$ of the unit sphere is, therefore, given by

$$d\sigma = \cos\phi' d\phi' d\lambda' , \quad (2-36)$$

and ψ , the spherical distance between \bar{P} and \bar{Q} , follows from the basic spherical triangle of Fig. 2.6:

$$\cos\psi = \sin\phi\sin\phi' + \cos\phi\cos\phi'\cos(\lambda' - \lambda) . \quad (2-37)$$

The function $S(\psi)$ has the form

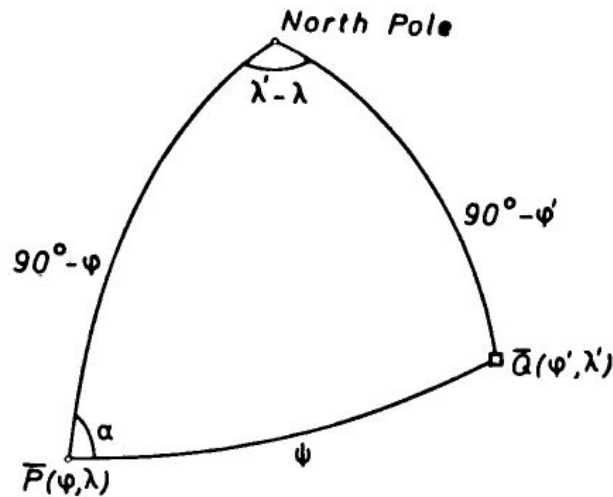


FIGURE 2.6. The basic spherical triangle.

$$S(\psi) = \left(\sin \frac{\psi}{2}\right)^{-1} - 6\sin \frac{\psi}{2} + 1 - 5\cos \psi - 3\cos \psi \ln \left(\sin \frac{\psi}{2} + \sin^2 \frac{\psi}{2}\right). \quad (2-38)$$

The geoidal height now results from (2-31):

$$N_P = \frac{R}{4\pi\gamma} \iint_{\sigma} \Delta g_Q S(\psi) d\sigma; \quad (2-39)$$

γ may be replaced by a global mean value such as 980 gal ($1 \text{ gal} = 10^{-2} \text{ m s}^{-2}$).

A differentiation of this formula according to (2-30) leads to *Vening Meinesz'* formula:

$$\begin{Bmatrix} \xi_P \\ \eta_P \end{Bmatrix} = \frac{1}{4\pi\gamma} \iint_{\sigma} \Delta g_Q \frac{dS}{d\psi} \begin{Bmatrix} \cos \alpha \\ \sin \alpha \end{Bmatrix} d\sigma, \quad (2-40)$$

where $dS/d\psi$ is the derivative of Stokes' function (2-38) and α is obtained from

$$\tan \alpha = \frac{\cos \phi' \sin(\lambda' - \lambda)}{\cos \phi \sin \phi' - \sin \phi \cos \phi' \cos(\lambda' - \lambda)}, \quad (2-41)$$

following from the spherical triangle of Fig. 2.6.

Both Stokes' formula and Vening Meinesz' formula presuppose the gravity anomaly Δg to be given at every point of the ellipsoid.

3. SPHERICAL HARMONICS

In this section we shall recall some well-known formulas for spherical harmonics for later reference; the notations follow (Heiskanen and Moritz, 1967), sections 1-8 through 1-15, 2-5, and 2-9.

Spherical coordinates r (radius vector), θ (polar distance), and λ (longitude) are related to rectangular coordinates x, y, z by

$$\begin{aligned} x &= r \sin \theta \cos \lambda , \\ y &= r \sin \theta \sin \lambda , \\ z &= r \cos \theta ; \end{aligned} \tag{3-1}$$

see Fig. 3.1.

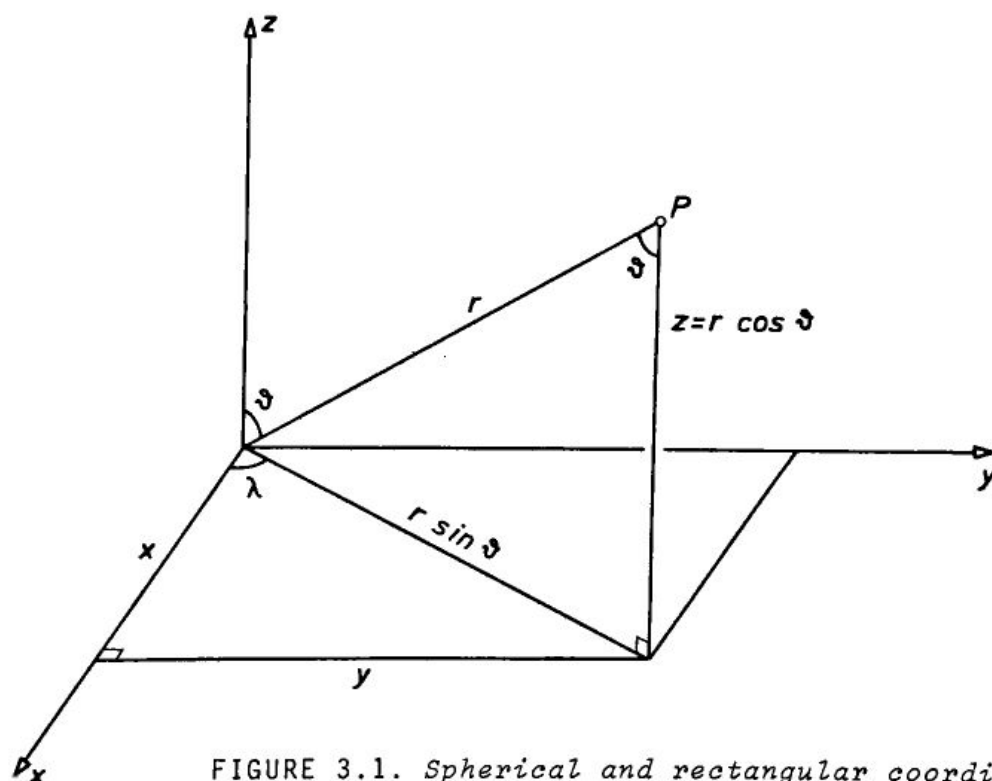


FIGURE 3.1. *Spherical and rectangular coordinates.*

If we express Laplace's equation $\Delta V = 0$ in spherical coordinates and try to solve it by a product of three functions, each of which depends on only *one* spherical coordinate:

$$V = f(r)g(\theta)h(\lambda) , \tag{3-2}$$

then the solutions are found to be

$$f(r) = r^n \quad \text{or} \quad f(r) = \frac{1}{r^{n+1}}, \quad (3-3)$$

$$g(\theta) = P_{nm}(\cos\theta), \quad (3-4)$$

$$h(\lambda) = \cos m\lambda \quad \text{or} \quad h(\lambda) = \sin m\lambda, \quad (3-5)$$

where

$$n = 0, 1, 2, 3, \dots$$

$$m = 0, 1, \dots, n.$$

n is called the *degree* and m , the *order* of the functions under consideration.

Thus, the dependence on r and on λ is simple: $f(r)$ is a positive or negative power of r , and $h(\lambda)$ is a sine or cosine of multiples of λ .

The functions $P_{nm}(\cos\theta)$ are less elementary. They are called Legendre functions and defined by (we put $\cos\theta = t$):

$$P_{nm}(t) = \frac{1}{2^n n!} (1-t^2)^{\frac{m}{2}} \frac{d^{n+m}}{dt^{n+m}} (t^2-1)^n. \quad (3-6)$$

An explicit expression is

$$P_{nm}(t) = 2^{-n} (1-t^2)^{\frac{m}{2}} \sum_{k=0}^{\nu} (-1)^k \frac{(2n-2k)!}{k!(n-k)!(n-m-2k)!} t^{n-m-2k}, \quad (3-7)$$

where ν is the greatest integer $\leq (n-m)/2$. The Legendre functions are thus polynomials in $t = \cos\theta$, multiplied by powers of $\sqrt{1-t^2} = \sin\theta$.

For $m = 0$ we have the *Legendre polynomials*

$$P_n(t) = P_{n0}(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n} (t^2-1)^n; \quad (3-8)$$

they are polynomials in t of degree n . For $m \neq 0$, the $P_{nm}(t)$ are called the *associated Legendre functions*.

The product of functions (3-4) and (3-5),

$$R_{nm}(\theta, \lambda) = P_{nm}(\cos\theta) \cos m\lambda, \quad (3-9)$$

$$S_{nm}(\theta, \lambda) = P_{nm}(\cos\theta) \sin m\lambda,$$

are Legendre surface harmonics, and the products of (3-3), (3-4), and (3-5),

$$\begin{aligned} r^n R_{nm}(\theta, \lambda), \quad r^n S_{nm}(\theta, \lambda), \\ r^{-(n+1)} R_{nm}(\theta, \lambda), \quad r^{-(n+1)} S_{nm}(\theta, \lambda), \end{aligned} \quad (3-10)$$

are the corresponding solid spherical harmonics ($m=0$: zonal; $m>0$: tesseral, $m=n$: sectorial). The functions (3-10), as well as their (finite or convergent infinite) linear combinations, are harmonic.

In particular, the series

$$V(r, \theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=0}^n \left[A_{nm} \frac{R_{nm}(\theta, \lambda)}{r^{n+1}} + B_{nm} \frac{S_{nm}(\theta, \lambda)}{r^{n+1}} \right], \quad (3-11)$$

or, equivalently,

$$V(r, \theta, \lambda) = \sum_{n=0}^{\infty} \frac{1}{r^{n+1}} \sum_{m=0}^n P_{nm}(\cos \theta) (A_{nm} \cos m\lambda + B_{nm} \sin m\lambda), \quad (3-12)$$

may be used for representing the earth's external gravitational potential, which is a harmonic function.

Since the first term, for $n=0$, is represented by GM/r , the series (3-11) or (3-12) may also be given the form, frequently used in satellite applications:

$$V = \frac{GM}{r} \left[1 - \sum_{n=2}^{\infty} \sum_{m=0}^n \left(\frac{a}{r} \right)^n P_{nm}(\cos \theta) (J_{nm} \cos m\lambda + K_{nm} \sin m\lambda) \right] \quad (3-13)$$

in which a is the semimajor axis of the earth (that is, of a best-fitting earth ellipsoid) and the coefficients J_{nm} and K_{nm} are, in a simple way, related to the coefficients A_{nm} and B_{nm} in (3-12). The advantage of the form (3-13) is that the coefficients are small dimensionless numbers. There is no term with $n=1$ if the origin is at the geocenter.

As an example we mention the case of the equipotential ellipsoid. In view of the rotational symmetry we have $K_{nm} = 0$ always and $J_{nm} = 0$ if $m \neq 0$. On putting $J_{n0} = J_n$ and noting (3-8), the expansion (3-13) thus reduces to

$$V = \frac{GM}{r} \left[1 - \sum_{n=2}^{\infty} J_n \left(\frac{a}{r} \right)^n P_n(\cos \theta) \right], \quad (3-14)$$

and the coefficients are given by

$$J_{2\nu} = (-1)^{\nu+1} \frac{3e^{2\nu}}{(2\nu+1)(2\nu+3)} \left(1 - \nu + 5\nu \frac{C-A}{ME^2} \right), \quad (3-15)$$

$$J_{2\nu+1} = 0,$$

using the notations of the preceding section and $\nu = 1, 2, 3, \dots$ (Heiskanen and Moritz, 1967, p.73).

Orthogonality relations. The integral over the unit sphere of the product of any two different functions R_{nm} or S_{nm} is zero:

$$\left. \begin{aligned} \iint_{\sigma} R_{nm}(\theta, \lambda) R_{sr}(\theta, \lambda) d\sigma &= 0 \\ \iint_{\sigma} S_{nm}(\theta, \lambda) S_{sr}(\theta, \lambda) d\sigma &= 0 \end{aligned} \right\} \text{ if } s \neq n \text{ or } r \neq m \text{ or both,} \quad (3-16)$$

$$\iint_{\sigma} R_{nm}(\theta, \lambda) S_{sr}(\theta, \lambda) d\sigma = 0 \quad \text{in any case.}$$

For the product of two equal functions we have

$$\begin{aligned} \iint_{\sigma} [R_{n0}(\theta, \lambda)]^2 d\sigma &= \frac{4\pi}{2n+1} = \kappa_{n0}, \\ \iint_{\sigma} [R_{nm}(\theta, \lambda)]^2 d\sigma &= \iint_{\sigma} [S_{nm}(\theta, \lambda)]^2 d\sigma = \\ &= \frac{2\pi}{2n+1} \frac{(n+m)!}{(n-m)!} = \kappa_{nm} \quad \text{if } m \neq 0. \end{aligned} \quad (3-17)$$

The integral over the unit sphere is expressed by

$$\iint_{\sigma} (\cdot) d\sigma = \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} (\cdot) \sin\theta d\theta d\lambda. \quad (3-18)$$

Let us now put $r = 1$ in (3-11) and write

$$V(1, \theta, \lambda) = f(\theta, \lambda), \quad (3-19)$$

so that

$$f(\theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=0}^n \left[A_{nm} R_{nm}(\theta, \lambda) + B_{nm} S_{nm}(\theta, \lambda) \right]. \quad (3-20)$$

22 General Background

We multiply $f(\theta, \lambda)$ by $R_{nm}(\theta, \lambda)$ or $S_{nm}(\theta, \lambda)$ and integrate over the unit sphere, taking into account (3-16) and (3-17). This determines the coefficients as

$$\begin{aligned} A_{nm} &= \frac{1}{\kappa_{nm}} \iint_{\sigma} f(\theta, \lambda) R_{nm}(\theta, \lambda) d\sigma, \\ B_{nm} &= \frac{1}{\kappa_{nm}} \iint_{\sigma} f(\theta, \lambda) S_{nm}(\theta, \lambda) d\sigma. \end{aligned} \quad (3-21)$$

Finally we introduce the *Laplace surface harmonics* $f_n(\theta, \lambda)$ of $f(\theta, \lambda)$, defined by

$$f_n(\theta, \lambda) = \sum_{m=0}^n \left[A_{nm} R_{nm}(\theta, \lambda) + B_{nm} S_{nm}(\theta, \lambda) \right], \quad (3-22)$$

and write

$$f(\theta, \lambda) = \sum_{n=0}^{\infty} f_n(\theta, \lambda). \quad (3-23)$$

Then the Laplace harmonic of degree n is given by the expression

$$f_n(\theta, \lambda) = \frac{2n+1}{4\pi} \int_{\lambda'=0}^{2\pi} \int_{\theta'=0}^{\pi} f(\theta', \lambda') P_n(\cos \psi) \sin \theta' d\theta' d\lambda', \quad (3-24)$$

which obviously is closely related to (3-21), ψ being the spherical distance between the points (θ, λ) and (θ', λ') :

$$\cos \psi = \cos \theta \cos \theta' + \sin \theta \sin \theta' \cos(\lambda' - \lambda). \quad (3-25)$$

Fully normalized harmonics. The "fully normalized" Legendre harmonics

$$\begin{aligned} \bar{R}_{nm}(\theta, \lambda) &= \sqrt{\frac{4\pi}{\kappa_{nm}}} R_{nm}(\theta, \lambda), \\ \bar{S}_{nm}(\theta, \lambda) &= \sqrt{\frac{4\pi}{\kappa_{nm}}} S_{nm}(\theta, \lambda), \end{aligned} \quad (3-26)$$

with κ_{nm} defined by (3-17), are not only orthogonal according to (3-16), but also normalized by

$$\frac{1}{4\pi} \iint_{\sigma} \bar{R}_{nm}^2 d\sigma = \frac{1}{4\pi} \iint_{\sigma} \bar{S}_{nm}^2 d\sigma = 1; \quad (3-27)$$

they form an *orthonormal system* of functions.

If the expansion (3-20) is written in terms of these functions:

$$f(\theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=0}^n \left[\bar{A}_{nm} \bar{R}_{nm}(\theta, \lambda) + \bar{B}_{nm} \bar{S}_{nm}(\theta, \lambda) \right] , \quad (3-28)$$

then the coefficients are given by

$$\begin{aligned} \bar{A}_{nm} &= \frac{1}{4\pi} \iint_{\sigma} f(\theta, \lambda) \bar{R}_{nm}(\theta, \lambda) d\sigma , \\ \bar{B}_{nm} &= \frac{1}{4\pi} \iint_{\sigma} f(\theta, \lambda) \bar{S}_{nm}(\theta, \lambda) d\sigma . \end{aligned} \quad (3-29)$$

Also the so-called *decomposition formula*, or *addition theorem*, for spherical harmonics takes a particularly simple form if expressed in fully normalized harmonics:

$$P_n(\cos\psi) = \frac{1}{2n+1} \sum_{m=0}^n \left[\bar{R}_{nm}(\theta, \lambda) \bar{R}_{nm}(\theta', \lambda') + \bar{S}_{nm}(\theta, \lambda) \bar{S}_{nm}(\theta', \lambda') \right] , \quad (3-30)$$

where $P_n(\cos\psi)$ is the usual Legendre polynomial (3-8) for the argument (3-25).

We finally mention the spherical-harmonic development of the *reciprocal distance*. Consider two points P and P' in space, having spherical co-ordinates

$$P(r, \theta, \lambda) \quad \text{and} \quad P'(r', \theta', \lambda') .$$

By applying the cosine theorem to the plane triangle OPP' , O being the origin $r = 0$, we find for the spatial distance $l = PP'$:

$$l = \sqrt{r^2 + r'^2 - 2rr' \cos\psi} , \quad (3-31)$$

where ψ , the angle between the radius vectors $r = OP$ and $r' = OP'$, is again given by (3-25). The reciprocal distance may now be expanded into the series

$$\frac{1}{l} = \sum_{n=0}^{\infty} \frac{r'^n}{r^{n+1}} P_n(\cos\psi) , \quad (3-32)$$

which converges (uniformly in ψ) for $r' < r$ since

$$|P_n(\cos\psi)| \leq 1 ; \quad (3-33)$$

it diverges for $r' > r$.

4. A FIRST LOOK AT HILBERT SPACE

The purpose of this and the following section is to provide an intuitive understanding of concepts such as operators and functionals which will be used throughout the present book. The treatment will be leisurely and simple, emphasizing analogies with vector and matrix calculus and not having too much concern about mathematical rigor. In fact, a certain familiarity with the basic terminology is, apart from a few exceptions, all that is needed for reading the book. If the reader wants to start research of his own, then he will wish to go further into the subject. A possible way to do so is first to read the lecture by Meissl (1975), then the one by Tscher-ning (1978a), and then to refer to one of the many good books, such as (Kantorovich and Akilov, 1964), (Kolmogorov and Fomin, 1970), or (Taylor, 1958).

The case of R^n . Consider first vectors and matrices in n -dimensional Euclidean space R^n . A vector x is an n -tuple of numbers

$$x = [x_1, x_2, \dots, x_n] ; \quad (4-1)$$

these vectors may also be interpreted as points of R^n . In this section, we shall write vectors as row vectors rather than column vectors, which is not quite consistent with standard matrix notation but convenient for generalization; furthermore, vectors and matrices will be symbolized by ordinary letters, without underlining.

The "size", or length or magnitude, of x is characterized by the *norm*

$$||x|| = (x_1^2 + x_2^2 + \dots + x_n^2)^{\frac{1}{2}} . \quad (4-2)$$

A transformation between a vector $x = [x_i]$ (this is an abbreviation for (4-1)) and another vector $y = [y_i]$ is given by

$$y = Ax , \quad (4-3)$$

where A is an $n \times n$ matrix with elements a_{ij} :

$$A = [a_{ij}] . \quad (4-4)$$

More explicitly, (4-3) may be written

$$y_i = \sum_{j=1}^n a_{ij} x_j \quad (i = 1, 2, \dots, n) . \quad (4-5)$$

Hilbert spaces. How can we generalize these relations? Let first $n \rightarrow \infty$, that is, the vector x is now an infinite sequence

$$x = [x_1, x_2, x_3, \dots] \quad (4-6)$$

for which the series

$$\|x\|^2 = \sum_{k=1}^{\infty} x_k^2 \quad (4-7)$$

converges; the norm $\|x\|$ defined in this way is a natural generalization of (4-2).

The set of all sequences (4-6) with finite norm forms a "space", which is an obvious generalization of n -dimensional Euclidian space formed by vectors (4-1). The space of such sequences is, of course, infinitely-dimensional; it is called the *Hilbert space of sequences* and denoted by l_2 .

A linear transformation in this Hilbert space is given by an analogue to (4-5):

$$y_i = \sum_{j=1}^{\infty} a_{ij} x_j \quad (i = 1, 2, 3, \dots) , \quad (4-8)$$

provided the sum, which here is an infinite series, converges. The matrix $A = [a_{ij}]$ is now an *infinite matrix*. It is also called a *linear operator* transforming the vector x into a vector y in a linear manner.

Another generalization is the *Hilbert space of square-integrable functions* L_2 . An intuitive approach is the following. The elements, or points, of R^n are vectors x_i as given by (4-1); the elements (or "points") of the present function space are functions, say $f(t)$, defined

on the interval $a \leq t \leq b$.¹ An analogy to (4-5) would then be

$$g(t) = \int_{u=a}^b A(t,u)f(u)du ; \quad (4-9)$$

there corresponds

$$\begin{array}{lll} \text{function } f(u) & \text{to} & \text{vector } x_j , \\ \text{argument } u & \text{to} & \text{index } j , \\ \text{integral } \int_{u=a}^b du & \text{to} & \text{sum } \sum_{j=1}^n . \end{array} \quad (4-10)$$

An expression of form (4-9) is called a *linear integral operator*; it transforms a function f into another function g . The function $A(t,u)$ is called the *kernel* of the operator.

Another example of a function space is provided by functions $f(\theta, \lambda)$ or $f(\phi, \lambda)$ defined on the unit sphere and satisfying certain conditions (such as continuity or square integrability). Linear integral operators are, for instance, Stokes' operator (2-39) and Vening Meinesz' operator (2-40); they transform the function Δg into the functions N, ξ, η , respectively.

However, not all linear operators are integral operators. The indefinite integral of a function $f(t)$,

$$F(t) = \int_a^t f(u)du , \quad (4-11)$$

may be expressed by an integral operator of form (4-9):

$$F(t) = \int_a^b A(t,u)f(u)du \quad (4-12)$$

¹The elements of L_2 are functions square integrable in the sense of Lebesgue. The Lebesgue integral is a rather advanced mathematical concept, which will not explicitly be used in this book. Therefore, the reader not interested in precise mathematical details may visualize our elements $f(t)$ simply as continuous functions for which the integrals which will occur in the sequel, especially the norm (4-30), are finite. In fact, some of the operators to follow are defined only for continuous or even differentiable functions.

with kernel

$$A(t,u) = \begin{cases} 1 & \text{if } u \geq t \\ 0 & \text{if } u < t \end{cases} . \quad (4-13)$$

The inverse operator, however, which is differentiation

$$f(t) = F'(t) , \quad (4-14)$$

cannot be expressed as an integral operator with an ordinary function as a kernel.

Nevertheless, the compact notation corresponding to (4-3),

$$g = Lf , \quad (4-15)$$

$$f = L^{-1}g , \quad (4-16)$$

L denoting a linear operator and L^{-1} the inverse operator (if it exists), can always be used; in our example, L would be integration and L^{-1} , differentiation.

The unit operator. Thus, what is the operator corresponding to the unit matrix? Acting on a function, it must leave the function unchanged:

$$If = f , \quad (4-17)$$

this can be considered as the definition of the *identity operator*, or *unit operator*, I . Can I be considered as an integral operator? Let us proceed by analogy to the case of R^n . There the equation

$$Ix = x , \quad (4-18)$$

I being the unit matrix, can be written in the form (4-5):

$$\sum_{j=1}^n \delta_{ij} x_j = x_i , \quad (4-19)$$

where the *Kronecker delta*

$$\delta_{ij} = \begin{cases} 1 & \text{if } j = i , \\ 0 & \text{if } j \neq i . \end{cases} \quad (4-20)$$

denotes the elements of the unit matrix. The continuous analogue of (4-19) would be

$$\int_{u=a}^b \delta(t,u)f(u)du = f(t) . \quad (4-21)$$

The kernel $\delta(t,u)$ would have to be zero for all $u \neq t$ (similar to $\delta_{ij} = 0$ for $j \neq i$), and for $u = t$ it should be infinite in such a way that the integral gives $f(t)$.

This behavior (zero or not) depends only on the difference $u - t$ (non-zero or not); therefore we may write

$$\delta(t,u) = \delta(u-t) , \quad (4-22)$$

where $\delta(x)$ would be a function of one variable only with

$$\delta(x) = 0 \quad \text{if} \quad x \neq 0 \quad (4-23)$$

and $\delta(0)$ infinite in such a way that

$$\int_{-e}^e \delta(x)dx = 1 \quad (4-24)$$

for arbitrarily small e . Then it is easy to see that (4-21) would hold for any continuous function defined in the (open) interval (a,b) . In fact, (4-21) becomes

$$\int_{u=a}^b \delta(u-t)f(u)du = \int_{u=t-e}^{t+e} \delta(u-t)f(u)du ,$$

in view of (4-22) and (4-23), and further

$$= \int_{t-e}^{t+e} \delta(u-t)f(t)du = f(t) \int_{t-e}^{t+e} \delta(u-t)du$$

because only the value of f at the point $u = t$ contributes to the integral. The last integral becomes (4-24), on substituting $x = u - t$, and has thus the value 1. This verifies (4-21).

Unfortunately, there is no ordinary function for which (4-23) and (4-24) hold. The symbol $\delta(x)$ denotes a "generalized function" or, in modern terminology, a *distribution* (cf. Kolmogorov and Fomin, 1970, p.124). Still, it has long been, and continues to be, customary to call $\delta(x)$ the *Dirac delta function*, or briefly, the *delta function*.

In this sense, the delta function may be considered as the kernel of the unit operator I although, strictly speaking, I is not an integral operator with an ordinary function as a kernel.

The delta function may be regarded, in a certain sense, as a limit of ordinary functions $\delta_n(x)$. We may, for instance, take

$$\delta_n(x) = \begin{cases} n, & -\frac{1}{2n} \leq x \leq \frac{1}{2n}, \\ 0 & \text{elsewhere,} \end{cases} \quad (4-25)$$

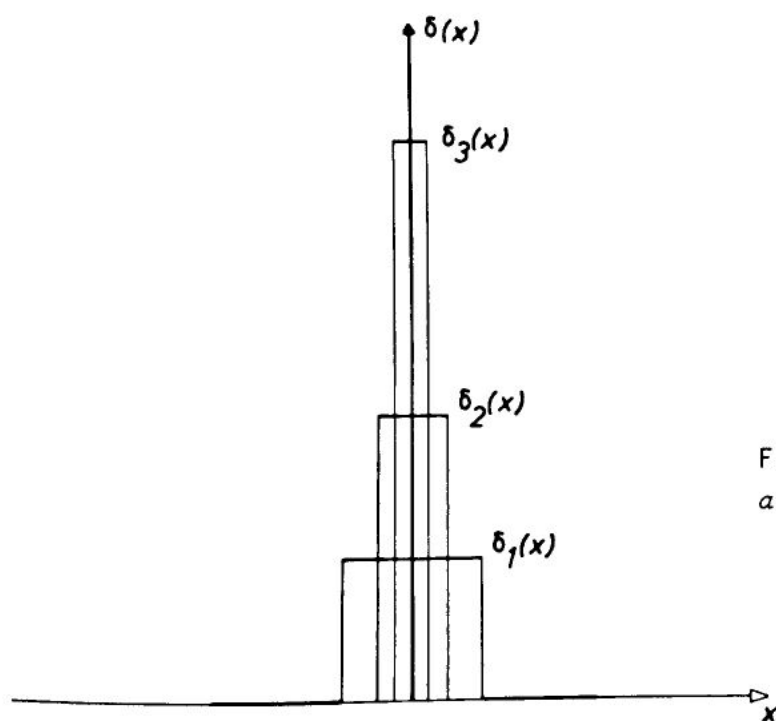


FIGURE 4.1. The delta function as a limit of ordinary functions.

see Fig. 4.1. This leads to an interpretation in terms of mean values; in fact, (4-21), with δ replaced by δ_n , gives the average value of the function f in the interval $(t-1/2n, t+1/2n)$. For $n \rightarrow \infty$, the length $1/n$ of this interval tends to zero, and the mean value tends to the point value $f(t)$, as it should be according to (4-21).

The functions (4-25) are discontinuous, but there are also continuous approximations to $\delta(x)$, e.g.

$$\delta_n(x) = \frac{n}{\sqrt{\pi}} e^{-n^2 x^2}.$$

The inner product. In R^n , the inner product of two vectors $x = [x_1, x_2, \dots, x_n]$ and $y = [y_1, y_2, \dots, y_n]$ is defined by

$$(x, y) = x_1 y_1 + x_2 y_2 + \dots + x_n y_n = \sum_{i=1}^n x_i y_i. \quad (4-26)$$

In vector calculus, the inner product is usually denoted by $x \cdot y$; in matrix notation, by $x^T y$; the notation (x, y) is customary in functional analysis.

The generalization to the Hilbert space l_2 is straightforward:

$$(x, y) = \sum_{i=1}^{\infty} x_i y_i. \quad (4-27)$$

In L_2 , the inner product between two functions f and g is naturally defined by

$$(f, g) = \int_{t=a}^b f(t)g(t)dt, \quad (4-28)$$

using the correspondence (4-10).

The inner product of a vector (or a function) with itself is nothing else than the square of the norm:

$$\|f\|^2 = (f, f), \quad (4-29)$$

which is in agreement with (4-2) and (4-7); for L_2 this means

$$\|f\|^2 = \int_a^b [f(t)]^2 dt. \quad (4-30)$$

Orthonormal systems. In R^n , a vector x may be represented in the form

$$x = x_1 e_1 + x_2 e_2 + \dots + x_n e_n, \quad (4-31)$$

where e_1, e_2, \dots, e_n are mutually orthogonal unit vectors. They satisfy

the orthonormality relations

$$(e_i, e_j) = \delta_{ij} , \quad (4-32)$$

which means that the norm of each vector e_i (the inner product of e_i with itself) is 1 (normalization), and that the inner product of two different vectors e_i and e_j is 0 (orthogonality). The vectors e_i form an *orthonormal base*. The component x_i is then simply the inner product of the vector x with the base vector e_i :

$$x_i = (x, e_i) . \quad (4-33)$$

This is immediately generalized to Hilbert space L_2 . Assume that there are base functions ϕ_i satisfying

$$(\phi_i, \phi_j) = \delta_{ij} \quad (4-34)$$

for $i, j = 1, 2, 3, \dots$, and that a given function f can be expanded into a series of such base functions

$$f = \sum_{i=1}^{\infty} f_i \phi_i ; \quad (4-35)$$

if such an expansion is possible, then the orthonormal system ϕ_i is called *complete*. For functions defined on the interval $[a, b]$, these relations mean

$$\int_a^b \phi_i(t) \phi_j(t) dt = \delta_{ij} , \quad (4-36)$$

$$f(t) = \sum_{i=1}^{\infty} f_i \phi_i(t) . \quad (4-37)$$

The coefficients f_i are then found by forming the inner product of f , as given by (4-28), with the base functions ϕ_i :

$$f_i = (f, \phi_i) . \quad (4-38)$$

In fact, denote the summation index by k and write (4-35) in the form

$$f = \sum_{k=1}^{\infty} f_k \phi_k ;$$

it is clear that summation indices and integration variables can be denoted by arbitrary letters. Then

$$(f, \phi_i) = \left(\sum_{k=1}^{\infty} f_k \phi_k, \phi_i \right) = \int_a^b \sum_{k=1}^{\infty} f_k \phi_k(t) \phi_i(t) dt .$$

Interchanging the order of integration and summation (this would have to be justified if we wished to proceed in a rigorous way) and using (4-34) we get

$$(f, \phi_i) = \sum_{k=1}^{\infty} f_k (\phi_k, \phi_i) = \sum_{k=1}^{\infty} f_k \delta_{ki} .$$

In this infinite sum, all terms vanish except the one with $k = i$, because $\delta_{ki} = 0$ for $k \neq i$. For $k = i$ we have $\delta_{ki} = 1$, so that there remains

$$(f, \phi_i) = f_i ,$$

which proves (4-38).

Isomorphism between L_2 and l_2 . Function space L_2 and sequence space l_2 are isomorphic; that is, there is a one-to-one correspondence between elements and relations in these two spaces. The principle is very simple: the expansion (4-37), for a given system of base functions $\phi_i(t)$, furnishes a correspondence between a function $f(t)$ and the infinite vector $[f_1, f_2, f_3, \dots]$ formed by the coefficients of the expansion of this function. It is less elementary to show that this correspondence is one-to-one: that to each function from L_2 there corresponds a sequence (an infinite vector) from l_2 and vice versa.

What is more, even the norms of two corresponding elements are equal:

$$\int_a^b [f(t)]^2 dt = \sum_1^{\infty} f_i^2 ,$$

(4-39)

and so are inner products:

$$\int_a^b f(t)g(t)dt = \sum_1^{\infty} f_1 g_1 . \quad (4-40)$$

We give a heuristic derivation of this relation, using (4-34) and (4-35):

$$\begin{aligned} \int_a^b f(t)g(t)dt &= (f, g) = \left(\sum_{i=1}^{\infty} f_i \phi_i, \sum_{j=1}^{\infty} g_j \phi_j \right) \\ &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f_i g_j (\phi_i, \phi_j) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} f_i g_j \delta_{ij} = \sum_{i=1}^{\infty} f_i g_i \end{aligned}$$

(of course, the interchange of sums and integrals would have to be justified).

To each operator A in L_2 there corresponds an infinite matrix in l_2 :

$$[a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots \\ a_{21} & a_{22} & a_{23} & \cdots \\ a_{31} & a_{32} & a_{33} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad (4-41)$$

the elements of which are defined by

$$a_{ij} = (A\phi_j, \phi_i). \quad (4-42)$$

Example 1. Let the Hilbert space L_2 be the set of functions square-integrable on the unit circle $0 \leq t < 2\pi$; that is, the functions are periodic with period 2π . The inner product is defined as

$$(f, g) = \int_0^{2\pi} f(t)g(t)dt . \quad (4-43)$$

A system of orthonormal base functions is

$$\begin{aligned}
\phi_1(t) &= \frac{1}{\sqrt{2\pi}} , \\
\phi_2(t) &= \frac{1}{\sqrt{\pi}} \cos t , & \phi_3(t) &= \frac{1}{\sqrt{\pi}} \sin t , \\
\phi_4(t) &= \frac{1}{\sqrt{\pi}} \cos 2t , & \phi_5(t) &= \frac{1}{\sqrt{\pi}} \sin 2t , \\
\phi_6(t) &= \frac{1}{\sqrt{\pi}} \cos 3t , & \phi_7(t) &= \frac{1}{\sqrt{\pi}} \sin 3t ,
\end{aligned}
\tag{4-44}$$

* * *

It is verified by direct integration that the orthonormality relations (4-34) are satisfied.

An expansion (4-35),

$$\begin{aligned}
f(t) = \frac{1}{\sqrt{\pi}} \left(\frac{f_1}{\sqrt{2}} + f_2 \cos t + f_4 \cos 2t + f_6 \cos 3t + \dots \right. \\
\left. + f_3 \sin t + f_5 \sin 2t + f_7 \sin 3t + \dots \right)
\end{aligned}
\tag{4-45}$$

is nothing else than the usual *Fourier series* (apart from constant factors ensuring that the system of base functions is not only orthogonal but also normalized).

Which infinite matrix $[d_{ij}]$ corresponds in this case to the differentiation operator D ? By definition,

$$Df(t) = f'(t) = \frac{df}{dt} . \tag{4-46}$$

We write

$$\begin{aligned}
f'(t) = g(t) = \sum_1^{\infty} g_i \phi_i(t) = \frac{1}{\sqrt{\pi}} \left(\frac{g_1}{\sqrt{2}} + g_2 \cos t + g_4 \cos 2t + \dots \right. \\
\left. + g_3 \sin t + g_5 \sin 2t + \dots \right)
\end{aligned}
\tag{4-47}$$

and compare this to the series obtained by termwise differentiation of (4-45):

$$\begin{aligned}
f'(t) = \frac{1}{\sqrt{\pi}} (-f_2 \sin t - 2f_4 \sin 2t - \dots \\
+ f_3 \cos t + 2f_5 \cos 2t + \dots) .
\end{aligned}$$

We find

$$\begin{aligned}
 g_1 &= 0, & g_2 &= f_3, & g_3 &= -f_2, \\
 g_4 &= 2f_5, & g_5 &= -2f_4, \\
 g_6 &= 3f_7, & g_7 &= -3f_6, \\
 &\cdot & \cdot & \cdot & \cdot
 \end{aligned}
 \tag{4-48}$$

This can be written as a matrix multiplication:

$$\begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ g_4 \\ g_5 \\ \cdot \\ \cdot \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & 1 & 0 & 0 & \cdot & \cdot \\ 0 & -1 & 0 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 2 & \cdot & \cdot \\ 0 & 0 & 0 & -2 & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \\ \cdot \\ \cdot \end{bmatrix}, \tag{4-49}$$

the square matrix being the matrix $[d_{ij}]$ corresponding to the operator \mathbf{I}

This may also be verified using (4-42). For instance, take the element d_{54} . We have

$$\phi_4 = \frac{1}{\sqrt{\pi}} \cos 2t,$$

$$D\phi_4 = \phi_4'(t) = -\frac{2}{\sqrt{\pi}} \sin 2t$$

$$\phi_5 = \frac{1}{\sqrt{\pi}} \sin 2t,$$

and hence, by (4-42),

$$d_{54} = (D\phi_4, \phi_5) = \int_0^{2\pi} \left(-\frac{2}{\sqrt{\pi}} \sin 2t\right) \frac{1}{\sqrt{\pi}} \sin 2t dt = -\frac{2}{\pi} \int_0^{2\pi} \sin 2t dt = -2,$$

in agreement with (4-49).

Example 2. Let our basic space be the set of functions square-integrable on the unit sphere ($0 \leq \theta \leq \pi$, $0 \leq \lambda < 2\pi$). The inner product is defined as

$$(f, g) = \frac{1}{4\pi} \iint_{\sigma} fg d\sigma, \quad (4-50)$$

which is the average product of the two functions over the sphere (the factor $1/4\pi$ does not change the basic fact that we have an integral of the product of the two functions; hence this definition of an inner product is permissible).

The fully normalized spherical harmonics (sec. 3) are orthonormal base functions in this space; for instance, (3-16) together with (3-27) corresponds to (4-34), (3-28) to (4-35), and (3-29) to (4-38).

Example 3. The unit operator I , defined by (4-17), corresponds to the infinite unit matrix

$$\begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot \\ 0 & 1 & 0 & \cdot & \cdot \\ 0 & 0 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (4-51)$$

in any system of base functions.

Linear functionals. An operator associates to a function (a vector) another function (another vector). A functional associates to a function (a vector) a real number.

In R^n , linear functionals are called linear forms. They associate to the n -vector x a number l by

$$l = \sum_{i=1}^n h_i x_i, \quad (4-52)$$

the coefficients h_i forming another vector which is a characteristic for the linear functional under consideration. In other terms, a linear functional in R^n may be represented as an inner product of two vectors h and x :

$$l = (h, x) \quad (4-53)$$

This representation as an inner product also holds for all functionals in a Hilbert space. For l_2 we have

$$l = \sum_{i=1}^{\infty} a_i x_i, \quad (4-54)$$

and for L_2 ,

$$l = \int_a^b h(t)f(t)dt, \quad (4-55)$$

with a function $h(u)$ characteristic for the linear functional. Generally we shall denote a linear functional of f by $L(f)$ and write

$$L(f) = l. \quad (4-56)$$

An example for a linear functional is the expression (4-38) for the coefficient f_i , which is a real number associated to the function f ; the base function ϕ_i takes the place of the characteristic function h in (4-55). In the space of functions defined on the sphere, Stokes' and Vening Meinesz' integrals (2-39) and (2-40) are examples of linear functionals if we consider N, ξ, η at one point only (for fixed ϕ, λ); then Stokes' integral (2-39) associates the value of N at this point to the function Δg . (If we regard the point (ϕ, λ) as variable and consider the function $N(\phi, \lambda)$, then Stokes' integral defines an operator, as we have seen above.)

A simple but important linear functional is the *evaluation functional*, or *delta functional*, associating to a function $f(t)$ its value at a given point $t = t_0$:

$$\delta_{t_0} f = f(t_0). \quad (4-57)$$

The name, evaluation functional, expresses the fact that the function is evaluated at the point t_0 , and the name, delta functional, arises from the possibility of writing (4-57) in the form

$$\delta_{t_0} f = f(t_0) = \int_{a=b}^b \delta(t-t_0)f(t)dt, \quad (4-58)$$

in view of (4-21), using the delta function (4-22).¹

The fundamental geodetic importance of linear functionals is due to the fact that all geodetic measurements depending on the gravity field may, after linearization, be considered as linear functionals of the anomalous potential T , as we shall see in Part B.

Norms of functionals and operators. The norm of a linear functional in Hilbert space, $\|L\|$, is simply the norm of the vector (or function) h ,

$$\|L\| = \|h\| . \quad (4-59)$$

the norm $\|h\|$ being given by (4-7) for l_2 and by (4-30) for L_2 . If the norm $\|L\|$ is finite, we speak of a *bounded* linear functional.

The norm of a linear operator L for which

$$g = Lf \quad (4-60)$$

according to (4-15), is the smallest number C (if it exists) for which

$$\|g\| \leq C\|f\| \quad (4-61)$$

for all elements f from the Hilbert space under consideration. We denote this norm by $\|L\|$:

$$\|L\| = C . \quad (4-62)$$

If $\|L\|$ is finite, then the operator L is called *bounded*; otherwise L is called *unbounded*.²

Examples. These important concepts will be illustrated by some examples.

¹ The evaluation functional (4-57) in L_2 is unbounded; otherwise a representation of form (4-58), but with an ordinary function instead of a delta function, would exist. In Hilbert spaces with a kernel function, the evaluation functional is a bounded linear functional; see sec. 24.

² Frequently the name, linear operator, is reserved for bounded linear operators, and similarly for functionals.

For the operator in R^3 , represented by the matrix

$$L = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix},$$

the length of the vector Lx is at most three times the length of the vector x . For instance, if

$$x = [1, 0, 0], \text{ then } Lx = [1, 0, 0],$$

but if

$$x = [0, 0, 1], \text{ then } Lx = [0, 0, 3]$$

(strictly speaking, we should write column vectors). Thus

$$\|L\| = 3.$$

Generally, in R^n , the norm $\|L\|$ is the maximum amount by which a vector x is stretched through the linear transformation Lx . For symmetric matrices, $\|L\|$ is the greatest eigenvalue.

In l_2 matters are similar. The infinite diagonal matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 1/2 & 0 & 0 & \dots \\ 0 & 0 & 1/3 & 0 & \dots \\ 0 & 0 & 0 & 1/4 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

has norm 1; its inverse

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots \\ 0 & 2 & 0 & 0 & \dots \\ 0 & 0 & 3 & 0 & \dots \\ 0 & 0 & 0 & 4 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

is unbounded.

The differential operator (4-46) is unbounded, for similar reasons; see the matrix in (4-49). (The norm of an operator in L_2 is the same as the norm of the associated infinite matrix.)

The unit operator I has norm 1 since $If = f$ for all f .

5. NORMED SPACES

In a Hilbert space, the norm of an element is defined in terms of the inner product of this element with itself:

$$\|f\|^2 = (f, f). \quad (5-1)$$

It is possible, however, to introduce a meaningful definition of the norm which does not depend on an inner product.

The space C of continuous functions. Consider the set of continuous functions $f(t)$ defined, e.g., on the interval $a \leq t \leq b$, and introduce a norm by

$$\|f\| = \max |f(t)|, \quad (5-2)$$

which is the greatest value which the absolute amount of $f(t)$ attains in the interval $[a, b]$; see Fig. 5.1. The set of continuous functions with the norm (5-2) forms the space C .

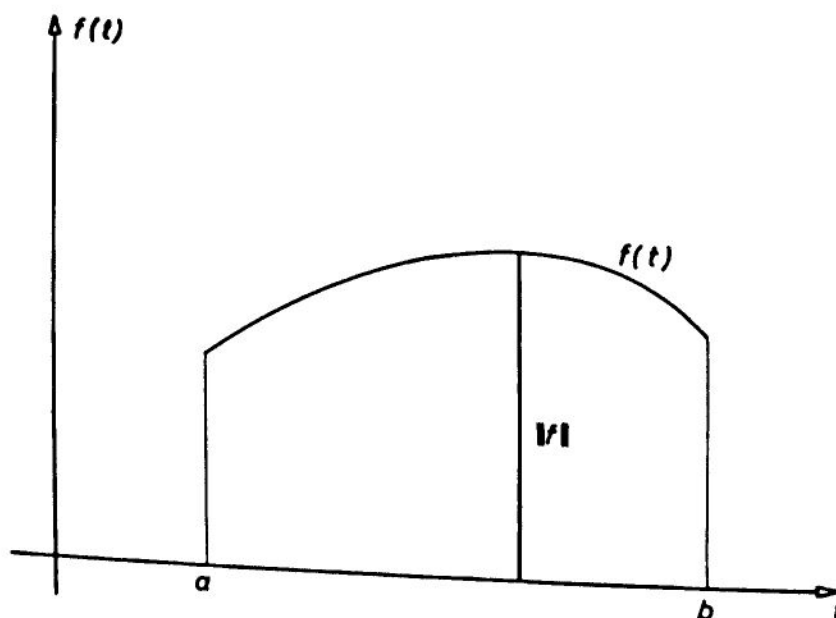


FIGURE 5.1. The norm in the space C .

Is this norm definition meaningful? Let us go back to the vector norm (4-2). Its purpose is to give a measure of the "length", or the "size", of the vector under consideration. Similarly, a norm of a function has to express, in a certain way, the size of this function. This can be done in different ways, depending on the purpose.

We have seen that the Hilbert norm (4-29) has been very appropriate for discussing orthogonal series such as Fourier series and expansions in spherical harmonics. The present norm (5-2) is closely related to the important concept of *uniform convergence*, which is found in any standard mathematical text (cf. Smirnow, 1967).

A sequence of continuous functions $f_n(t)$, $n = 1, 2, 3, \dots$, defined on the interval $[a, b]$, is said to converge uniformly to a continuous function $f(t)$ if the difference

$$f(t) - f_n(t)$$

tends uniformly to zero, that is, if the maximum absolute amount of this difference, for any t in $[a, b]$,

$$\max |f(t) - f_n(t)|$$

tends to zero if $n \rightarrow \infty$. In terms of the norm (5-2) this simply means

$$\|f - f_n\| \rightarrow 0 \quad \text{if } n \rightarrow \infty. \quad (5-3)$$

Thus, convergence in the norm (5-2) is nothing else than uniform convergence in the classical sense. Therefore, the norm (5-2) is also called the uniform norm.

By means of convergence in the norm it is also possible to introduce the important concept of *completeness*.

Assume that a sequence of functions $f_n(t)$ converges uniformly to a function $f(t)$, all functions being elements of C , so that (5-3) holds. Then the sequence of $f_n(t)$ also satisfies the *Cauchy criterion* for uniform convergence: given any $\epsilon > 0$, there is an integer N such that

$$\|f_n - f_m\| < \epsilon \quad \text{for all } n, m > N. \quad (5-4)$$

A sequence of f_n satisfying (5-4) is called a *Cauchy sequence*.

It is now well known that the limit of a uniformly convergent sequence of continuous functions is itself a continuous function. This means that

every Cauchy sequence in the space C converges to an element f of C , which precisely is meant by saying that the space C is *complete*.

In an incomplete space, not every Cauchy sequence of elements has a limit which is an element of the space.

Completeness in R . The simplest illustration for completeness and incompleteness is furnished by the real number line $R: -\infty < x < \infty$. For each real number there exists a Cauchy sequence of rational numbers. Take, for instance, the irrational number $\sqrt{2}$; then the sequence

$$\begin{aligned} x_1 &= 1.4, \\ x_2 &= 1.41, \\ x_3 &= 1.414, \\ x_4 &= 1.4142, \\ x_5 &= 1.41421, \\ &\vdots \end{aligned} \tag{5-5}$$

obtained by truncating the infinite decimal fraction for $\sqrt{2}$ after the n -th decimal, forms a Cauchy sequence for $\sqrt{2}$.

Consider now the set Q formed only of the rational numbers on R (the irrational numbers are removed). Then the Cauchy sequence x_n consists of elements of Q , but this sequence does not converge to an element of Q since $\sqrt{2}$ is not a rational number.

It is clear that R is a normed space with norm

$$\|x\| = |x|,$$

which is the absolute amount of the number x . But the set Q , with the same norm, also satisfies the conditions (5-9) of a normed space to be given below: Q is an incomplete normed space. On the other hand, R is complete since every Cauchy sequence of real (rational and irrational) numbers converges to a real number (in our example, $\sqrt{2}$).

Denseness. We have just seen that real numbers (such as $\sqrt{2}$) can be approximated arbitrarily well by rational numbers: for any given real number x there is a rational number x_n such that

$$|x - x_n| < \epsilon \tag{5-6}$$

for any (arbitrarily small) positive number ϵ . We say that the rational numbers are *dense* in the real numbers, or that Q is a dense subset of R .

In an obvious generalization we say that a set of elements f_n ($n = 1, 2, 3, \dots$) is dense in some space if any element f of the space can be approximated arbitrarily well by some f_n , in the sense that

$$\|f - f_n\| < \epsilon \quad (5-7)$$

for any positive ϵ .

For example, the set of polynomials is dense in the function space C since any continuous function can be arbitrarily well approximated uniformly by polynomials, that is, in the sense of the metric (5-2). This is the *Theorem of Weierstrass*, well known from approximation theory (cf. Davis, 1975, p.108).

The density theorem most important in physical geodesy is Runge's theorem, to be discussed in secs. 7 and 8.

At this point we are able to render the notion of abstract spaces somewhat more precise. In keeping with the general character of this introductory treatment we shall not give a full axiomatic characterization of such space (which can e.g. be found in (Tscherning, 1978a)) but point out some essential features.

A *linear space*, or linear vector space, consists of elements f, g, h, \dots such that sums, differences, multiples, and generally linear combinations of these elements,

$$f + g, \quad f - g, \quad \alpha f, \quad \alpha_1 f + \alpha_2 g + \alpha_3 h, \dots, \quad (5-8)$$

are also elements of the space¹; $\alpha, \alpha_1, \alpha_2$, etc., here and in the sequel, denote arbitrary real numbers. Addition, subtraction, and multiplication by a real number "does not lead out of the space".

It is clear that this property is satisfied by vectors, but also, e.g., by harmonic functions: the linear combination of harmonic functions is also harmonic, in view of the linearity of Laplace's equation.

A *normed space* is a linear space for the elements f of which there is defined a norm possessing the following properties:

¹ If the elements of the space are ordinary vectors or functions, then addition and multiplication by a number α are understood in the usual way; for abstract elements, these operations can be introduced axiomatically; cf. (Tscherning, 1978a, p.159).

$$\begin{aligned}
\|f\| &\geq 0 && \text{(positivity),} \\
\|\alpha f\| &= |\alpha| \cdot \|f\| && \text{(homogeneity),} \\
\|f+g\| &\leq \|f\| + \|g\| && \text{(triangle inequality),} \\
\|f\| &= 0 \text{ if and only if } f = 0.
\end{aligned}
\tag{5-9}$$

These conditions are easily seen to be satisfied by the usual vector norm (4-2), but they can also be verified by Hilbert space norms such as (4-30) and by the norm (5-2) in the space C .

An *inner-product space* is a linear space if an inner product is defined which possesses the following properties

$$\begin{aligned}
(f, g) &\text{ is a real number,} \\
(f, g) &= (g, f) && \text{(symmetry),} \\
(f_1 + f_2, g) &= (f_1, g) + (f_2, g) && \text{(distributivity),} \\
(\alpha f, g) &= \alpha(f, g) && \text{(homogeneity),} \\
(f, f) &\geq 0 \text{ and zero if and only if } f = 0.
\end{aligned}
\tag{5-10}$$

An inner-product space is always a normed space if the norm is defined by (5-1) but the opposite does not hold: the space C is a normed space but an inner product related to the norm is not defined on it.

There are complete and incomplete normed spaces, as we have seen above. A complete normed space is called a *Banach space*, a complete inner-product space is a *Euclidean space* if its dimension is finite and a *Hilbert space* if its dimension is infinite. Finite dimension means that all elements of the space can be represented as linear combinations of n "base elements", for instance, of the elements e_i in (4-31); for Hilbert space the number of base elements is infinite; cf. (4-35).¹

The reader who meets these concepts and definitions for the first time may ask for what use they serve. The definitions are introduced in such a way as to extend simple and well-known properties (of numbers, vectors, etc.) to more general cases (such as functions), for instance, the property of completeness. It also turns out that simple relations holding in ordinary three-dimensional space can be generalized in a natural way to abstract spaces, for instance to function spaces. This helps to give an in-

¹ The terminology is not uniform: sometimes all complete inner-product spaces are called Hilbert spaces, sometimes all are called Euclidean spaces. Incomplete inner-product spaces are also given the name of *pre-Hilbert spaces*.

tuitive "geometric" flavor to abstract situations and provides a convenient vocabulary. The exhibition of a common structure in apparently quite different contexts (e.g., the orthogonality of vectors and the orthogonality of functions) greatly contributes to our understanding by unifying, ordering, and simplifying it.

Linear functionals in a normed space. A linear functional $L(f)$ associates to an element f a real number l :

$$L(f) = l, \quad (5-11)$$

and the linearity condition holds:

$$L(\alpha_1 f + \alpha_2 g) = \alpha_1 L(f) + \alpha_2 L(g). \quad (5-12)$$

A linear functional is *bounded* provided there exists a constant C such that

$$|L(f)| \leq C \|f\| \quad (5-13)$$

for all elements of the space. The smallest number C for which (5-13) holds is called the *norm* of the functional L and denoted by $\|L\|$.

The definition of the norm of a functional by (5-13) corresponds to the norm definition for a linear operator by (4-61). For an inner-product space, the functional

$$L(f) = (h, f) \quad (5-14)$$

according to (4-53) satisfies the defining condition (5-12), and the norm definition (4-59) can be shown to follow from (5-13).

Functionals on the space C . Any bounded linear functional in the space of functions continuous on the interval $[a, b]$ has the form

$$L(f) = \int_a^b f(t) dv(t) \quad (5-15)$$

or briefly,

$$L(f) = \int_a^b f dv \quad (5-16)$$

which is a Stieltjes integral, $v(t)$ being a so-called function of bounded variation. This is the *Riesz' representation theorem*; cf. (Kolmogorov and Fomin, 1970, § 36.6).

We shall need the analogue of this theorem for functions $f(x)$ continuous on a compact (i.e., closed and bounded) set K in three-dimensional space R^3 , $x = [x_1, x_2, x_3]$ denoting a point in R^3 ; the norm $\|f\|$ is the maximum of $|f|$ on K . We have

$$L(f) = \iiint_K f(x) dv(x) ; \quad (5-17)$$

cf. (Kantorovich and Akilov, 1964, ch. VI, § 4).

Without going into mathematical details, we give a physical interpretation, which is readily understood and important in potential theory. Consider first a functional

$$L^+(f) = \iiint_K f(x) d\mu(x) \quad (5-18)$$

in which $d\mu(x)$ represents a *mass element* in the physical sense, which is always positive. The masses can be distributed continuously over K or over part of K , but they can also be concentrated at points, on curves, and on surfaces, as illustrated in Fig. 5.2. For example, the mass element $d\mu = d\mu(x)$ symbolically represented in this figure contains the point mass m_2 situated at the point x , as well as part of a continuous distribution and of a surface layer on S . If P denotes a point outside K and l is the distance of P from $d\mu$, then

$$V(P) = G \iiint_K \frac{d\mu}{l} \quad (5-19)$$

is nothing else than the gravitational potential generated at P by the masses situated within K . It is clearly a linear functional of type (5-18), with

$$f(x) = \frac{G}{l} ,$$

G being the gravitational constant.

Let $d\mu_1(x)$ and $d\mu_2(x)$ denote two of such positive mass distributions within K . Their difference

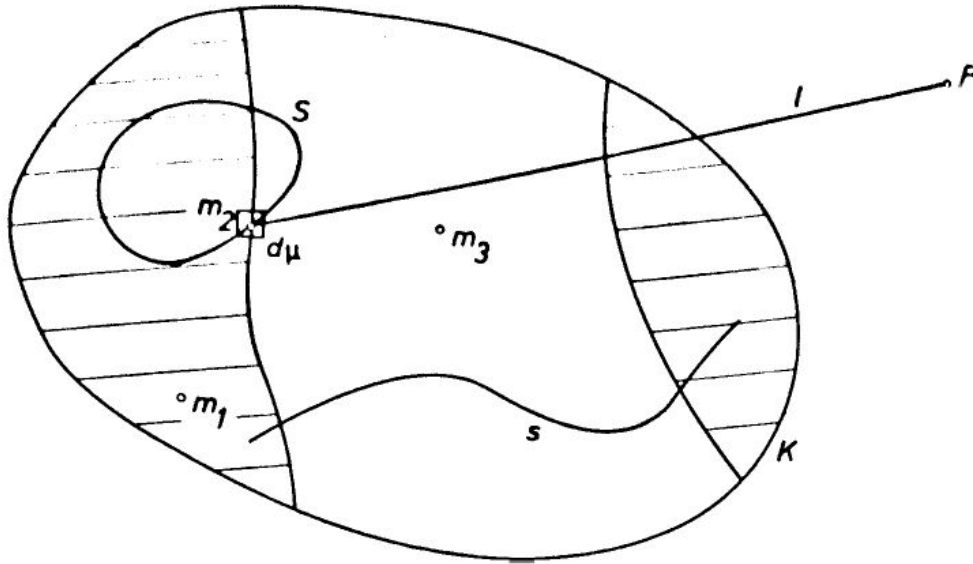


FIGURE 5.2. A mass distribution in the compact set K , consisting of masses continuously distributed in the two shaded regions, as well as concentrated at the points m_1, m_2, m_3 , on the line s and on the closed surface S .

$$dv(x) = d\mu_1(x) - d\mu_2(x) \quad (5-20)$$

may be positive or negative. It has the physical character of a *charge* (in the sense of an electric charge); in mathematical terms, $v(x)$ is a (*signed*) *measure*. It may be shown that $dv(x)$ in (5-17) always has the form (5-20): any bounded linear functional $L(f)$ on the space C can be represented as an integral of f with respect to a certain charge distributed in K .

Similarly, (5-15) may be interpreted as an integral of f with respect to charges distributed on the segment $[a, b]$.

Nonlinear operators. A fundamental notion in modern mathematics is the concept of a *mapping*. Think of a geographical map. To each point of a certain part of the earth's surface it associates a point in the plane. But also a continuous function $f(t)$ defined on the interval $[a, b]$ may be

considered as such a mapping: to each point t_0 in the interval $[a, b]$ it associates a real number $f(t_0)$, which is a point of the space R of real numbers α , $-\infty < \alpha < \infty$; it is a mapping of the interval into R . We may express this by writing

$$f: [a, b] \rightarrow R. \quad (5-21)$$

Let us illustrate this concept by other examples. Take the gravitational potential

$$V = V(x_1, x_2, x_3), \quad (5-22)$$

which is (1-1) with $x_1=x$, $x_2=y$, $x_3=z$. It is defined in the whole space R^3 . To each point $[x_1, x_2, x_3]$ there corresponds a real number, which is the value of the function V at this point, this real number is an element of R . We may thus consider the potential V a mapping of R^3 into R , symbolically

$$V: R^3 \rightarrow R. \quad (5-23)$$

The gravitational vector $\text{grad } V$ is a mapping

$$\text{grad } V: R^3 \rightarrow R^3, \quad (5-24)$$

because it associates to each spatial point $[x_1, x_2, x_3]$ the three components of $\text{grad } V$, say $[a_1, a_2, a_3]$, which may also be considered as a point in R^3 .

A spatial curve r connecting two points A and B in space may be regarded as a mapping

$$r: [a, b] \rightarrow R^3. \quad (5-25)$$

In fact, take the parameter representation

$$x_1 = x_1(t), \quad x_2 = x_2(t), \quad x_3 = x_3(t),$$

and let to points A and B correspond the parameter values $t = a$ and $t = b$, respectively. Then to each parameter value in the interval $[a, b]$ there corresponds a point $[x_1, x_2, x_3]$ of the curve, which is precisely the meaning of (5-25).

Similarly with surfaces. For instance, the unit sphere σ is a mapping of the rectangle $[0 \leq \theta \leq \pi, 0 \leq \lambda < 2\pi]$ (θ and λ are here considered as rectangular coordinates) into R^3 , given by

$$x_1 = \sin\theta \cos\lambda, \quad x_2 = \sin\theta \sin\lambda, \quad x_3 = \cos\theta.$$

Consider now an arbitrary set X of elements x ; we write

$$x \in X \tag{5-26}$$

to denote that x is an element of X . For instance,

$$[x_1, x_2, x_3] \in R^3, \quad f(t) \in C.$$

Consider another set Y of elements y , such that $y \in Y$. Let there be given a rule which associates to each $x \in X$ one element $y \in Y$, symbolically

$$y = F(x) \quad \text{or} \quad y = Fx. \tag{5-27}$$

Such a rule is called a *mapping* of the set X into the set Y ; one writes

$$F: X \rightarrow Y. \tag{5-28}$$

Such a *mapping* F is also called an *operator* or a *function* (in a generalized sense).

The set X , on which F is defined, is called the *domain* of F ; the set of all those elements of Y which are images Fx , is the *range* of F . The domain of F , by definition, is the whole set X , but the range is, in general, only a subset of Y (the curve r in (5-25) does not fill the whole space R^3 !).

This concept is very general and useful. Clearly, (5-21), (5-23), and (5-25) are special cases of such a mapping, but also the preceding and the present section abound in examples. The linear transformation (4-3) is a mapping $R^n \rightarrow R^n$, the linear integral operator (4-9) is a mapping

$$A: L_2 \rightarrow L_2,$$

eq. (4-37) is a mapping $l_2 \rightarrow L_2$ (to an infinite vector $[f_1, f_2, \dots] \in l_2$

there is associated a function $f \in L_2$), and a linear functional such as (4-55) or (5-15) is a mapping

$$L: L_2 \rightarrow R \text{ or } C \rightarrow R.$$

Generally, a functional is a mapping

$$X \rightarrow R,$$

since it associates to an element $x \in X$ a real number $\in R$. A functional may be considered as a special case of an operator, for the case that $Y = R$.

It may be pointed out that X need not be a whole normed space. For instance, in (5-25), X is the interval $[a,b]$, which is a subset of R . Another example is the differentiation operator $D = d/dt$ in the space C . The domain of D is not the whole space C , but only a subset consisting of the differentiable functions $\in C$.

An operator (mapping, function) of type (5-28) is linear if

$$F(\alpha_1 x_1 + \alpha_2 x_2) = \alpha_1 F(x_1) + \alpha_2 F(x_2) \quad (5-29)$$

for all elements $x_1 \in X$ and $\alpha_1 \in R$.

Important operators are nonlinear. For instance, the norm is a nonlinear functional: (5-2) is a nonlinear function $C \rightarrow R$, and similarly for (5-1). Curves and surfaces are, in general, nonlinear mappings.

With nonlinear functions defined in normed spaces (i.e., nonlinear operators), a differential and integral calculus can be developed which is an elegant generalization of differentiation and integration in R^n ; cf. (Dieudonné, 1960) and (Loomis and Sternberg, 1968).

The concepts of linear and nonlinear operators and functionals will be frequently used in the sequel.

6. CONVERGENCE OF SPHERICAL HARMONICS I

The convergence problem for spherical harmonics is already an advanced topic. It is far from elementary and, to the author's knowledge, has not yet been fully solved from a theoretical point of view. A solution relevant for practical purposes is given by an application of Runge's theorem to be discussed in sec.7. The present development closely follows (Moritz, 1978e).

A sufficient condition for convergence is easily found. We write (1-1) in the form

$$V = V(P) = G \iiint_{\tau} \frac{dM}{l}, \quad (6-1)$$

where

$$dM = \rho dv$$

is the mass element and the integral is extended over the earth's body τ . Using the notations of Fig. 6.1 and formula (3-32), we obtain on substitution into the integral above and termwise integration:

$$V = \sum_{n=0}^{\infty} \frac{f_n(\theta, \lambda)}{r^{n+1}} \quad (6-2)$$

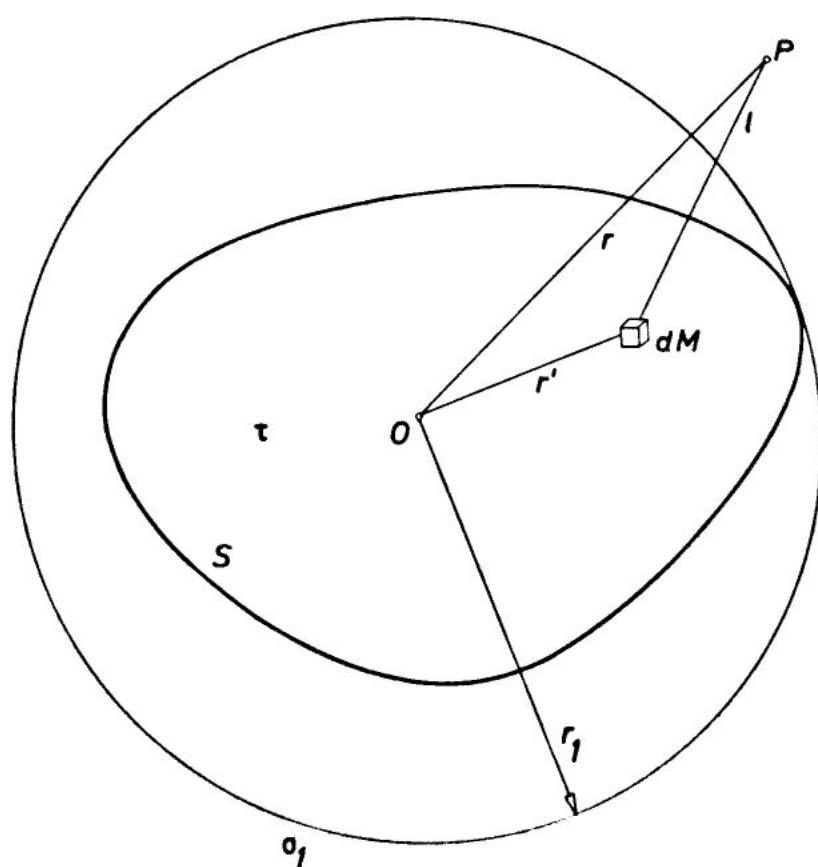


FIGURE 6.1. Convergence outside the sphere σ_1 .

where

$$f_n(\theta, \lambda) = G \iiint_{\tau} r'^n P_n(\cos \psi) dM \quad (6-3)$$

is readily seen to be identical to (3-22). Thus, (6-2) is the spherical-harmonic series for the external gravitational potential in the most condensed notation.

This termwise integration (interchange of summation and integration) is permissible as long as the series (6-2) converges uniformly (Kellogg, 1929, p.143). This is the case if $r' < r$. Thus the convergence can be assured wherever $r > r_1$, where r_1 is the maximum value which r' can attain, that is, for all points outside the sphere σ_1 of radius r_1 . In other words, the spherical-harmonic series for the external potential converges at all points outside a sphere σ_1 which, somewhat loosely speaking, is the smallest sphere (around the point taken as origin $r = 0$) that contains the body τ completely in its interior.

So far, all is standard. For the convergence of the spherical-harmonic series it is thus sufficient that the point P , at which V is considered, lies outside the sphere σ_1 . Is this condition also necessary? In other words, do we have divergence at all points inside and on σ_1 ? If P lies inside σ_1 , then the series (3-32) partly (for points with $r' > r$) diverges, and it would be tempting to conclude from this that (6-2) diverges inside σ_1 . However, such a conclusion is not correct because it is known that formal operations with divergent series may very well lead to convergent results; this was already remarked by Helmert (1884, p.70).

A convergent series. In fact, there are simple examples of spherical-harmonic series for which the region of convergence extends well within the sphere σ_1 . Such series were given by Helmert (1884, p.125), Levallois (1973), and others. We shall show that also the spherical-harmonic expansion for the external potential of the level ellipsoid converges at the surface of the ellipsoid and even well below.

This series is given by (3-14) and (3-15); we write it in the form

$$V = 3GM \sum_{n=0}^{\infty} \frac{(-E^2)^n}{(2n+1)(2n+3)} \left(1 - n + 5n \frac{C-A}{ME^2}\right) \frac{P_{2n}(\cos \theta)}{r^{2n+1}}, \quad (6-4)$$

E being again the linear excentricity (Fig. 6.2).

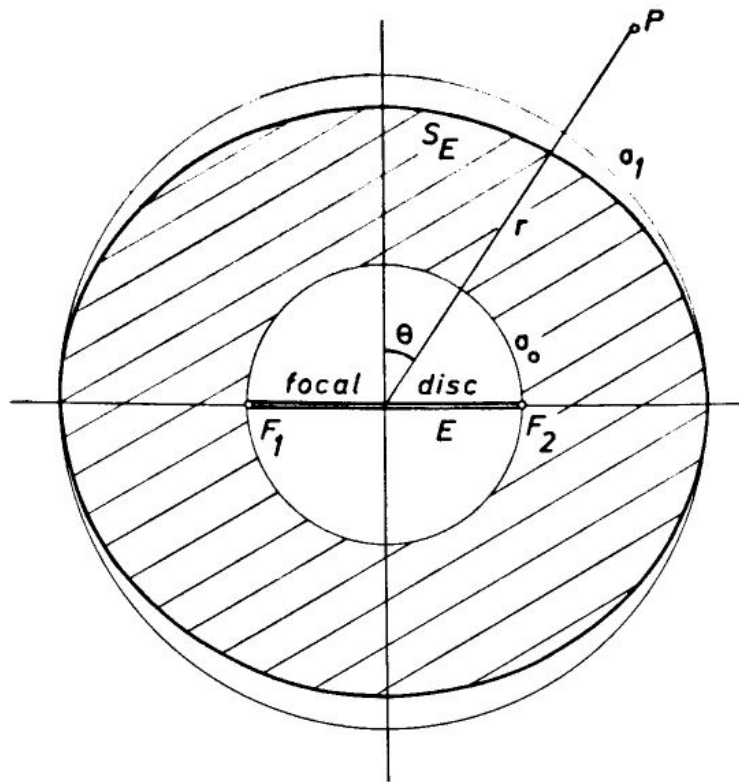


FIGURE 6.2. Convergence for the ellipsoid.

A majorant of the series (6-4), disregarding the factor $3GM/r$, is

$$\sum_{n=0}^{\infty} \frac{1 + \alpha n}{(2n+1)(2n+3)} x^n, \quad (6-5)$$

where we have put

$$\alpha = \left| 5 \frac{C-A}{ME^2} - 1 \right| \quad (6-6)$$

and

$$x = \frac{E^2}{r^2}; \quad (6-7)$$

again we have used (3-33). The quotient criterium shows at once that the series (6-5), and a fortiori the series (6-4), is convergent for $x < 1$, that is, for $r > E$. Hence the series (6-4) converges down to the sphere σ_0 which contains the two focal points (Fig. 6.2). If the sphere σ_0 lies inside the ellipsoid, which is certainly the case if the ellipsoidal flattening is as small as the flattening of the earth, then the series converges at the entire surface S_E of the ellipsoid.

It even converges in the interior of the ellipsoid in the region between the surfaces S_E and σ_0 (shaded in Fig. 6.2). It would, however, be wrong to assume that it represents the interior gravitational potential in this region. This is impossible because a spherical harmonic series always represents a *harmonic function*, that is, a solution of Laplace's equation $\Delta V = 0$, whereas the interior potential satisfies Poisson's equation (1-10). If the ellipsoid has a continuous distribution of density ρ that is positive in the whole interior of S_E , then Laplace's equation is incompatible with Poisson's equation, so that the series (6-4) represents a function V different from the interior potential V . The harmonic function represented by (6-4) in the interior of S_E is the so-called *analytical continuation* of the external potential into the interior of the body.

A closed expression corresponding to the series (6-4) is provided by eq. (2-59) of (Heiskanen and Moritz, 1967, p.66), which expresses the exterior potential together with its analytical continuation into the interior of the ellipsoid. It is easily seen that the function given by this formula is singular at the focal points F_1 and F_2 and on the whole "focal disc" represented in Fig. 6.2 by the segment joining F_1 and F_2 ; everywhere else it is regular. The sphere σ_0 is the smallest sphere that contains all singularities in its interior or on its surface, or the smallest sphere outside of which the harmonic function is everywhere regular.

Analytical continuation and convergence. In this example, the convergence behavior is determined by the behavior of the analytical continuation \hat{V} of the external potential, in particular with respect to the singularities of \hat{V} . It is readily seen that this fact is not restricted to the ellipsoid but is completely general.

In fact, consider (Fig. 6.3) the largest sphere σ_2 around the origin O that lies completely within the earth's surface S (more precisely, that does not include points of the space outside S). Assume that the external potential V can be regularly continued down to the surface σ_2 , so that V together with \hat{V} constitutes a harmonic function regular everywhere outside and on σ_2 . Then, it follows from a standard theorem of potential theory (solution of the exterior Dirichlet problem for the sphere

in terms of spherical harmonics, cf. Smirnow, 1964b, § 136) that this harmonic function can be expanded into a spherical-harmonic series that converges everywhere outside and on σ_2 .

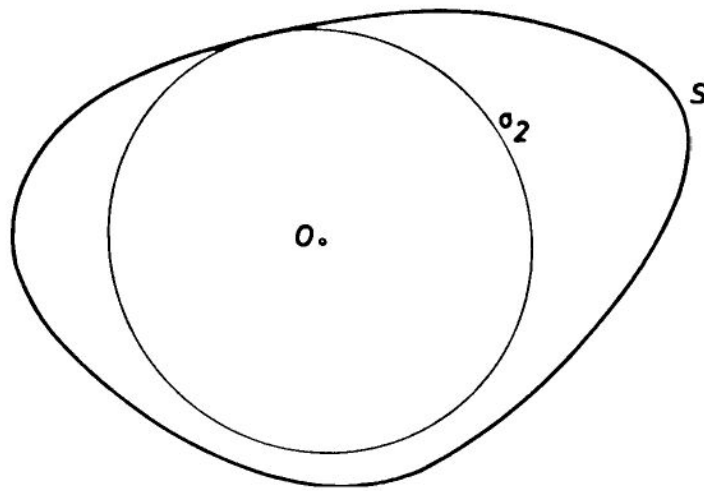


FIGURE 6.3. Analytical continuation and convergence.

Thus, in order to investigate the convergence of the spherical-harmonic series for the external gravitational potential at the earth's surface, we must analytically continue it downward to the sphere σ_2 . If this analytical continuation is regular everywhere outside and on σ_2 , then the series converges everywhere outside and on σ_2 . A fortiori, it then converges everywhere outside and on the earth's surface S and represents there the external gravitational potential.

Is there a sphere of convergence? This reasoning can be extended even further. Denote by σ_0 the smallest sphere about O outside of which the external potential together with its analytical continuation is regular (Fig. 6.4). The existence of such a sphere is clear; it will be called *limit sphere*. If the external potential cannot be analytically continued into the interior of the earth, then σ_0 coincides with σ_1 ; otherwise it lies inside σ_1 . An example for a limit sphere σ_0 is the sphere containing the focal points in the case of the level ellipsoid (Fig. 6.2).

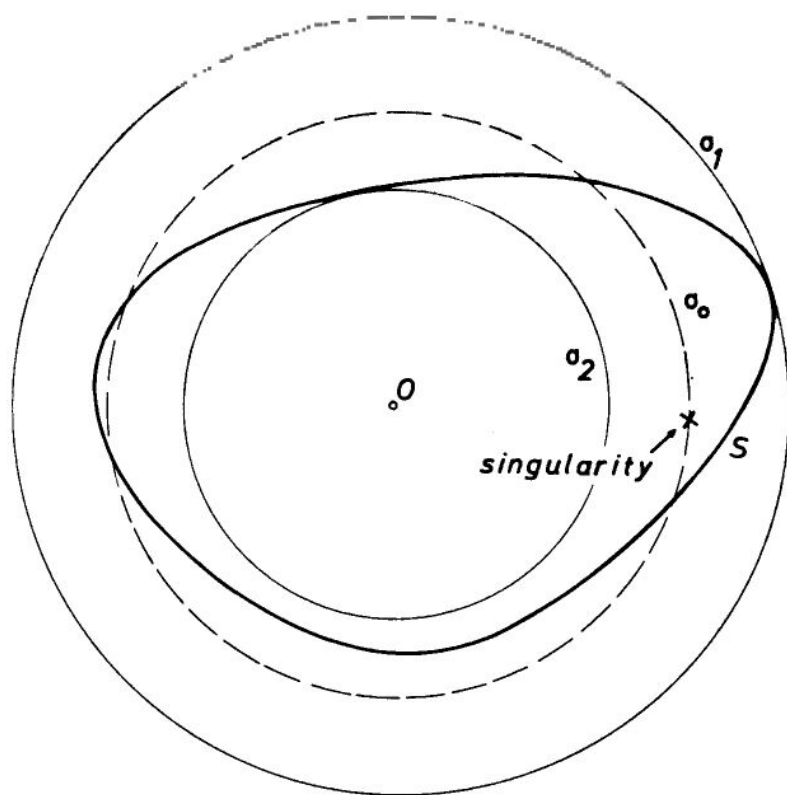


FIGURE 6.4. The limit sphere σ_0 .

The spherical-harmonic expansion for the potential will converge at all points outside σ_0 . To see this, take a concentric sphere σ slightly larger than σ_0 and apply the reasoning of the preceding section to σ instead of σ_2 : the spherical-harmonic series will converge everywhere outside and on σ . Since σ can be arbitrarily close to σ_0 , our assertion follows.

In terms of the limit sphere σ_0 we may obviously state that the spherical-harmonic series for the potential will converge on the earth's surface S if σ_0 is inside σ_2 .

It is also clear that the limit sphere σ_0 may be defined as the smallest sphere about 0 that contains all singularities of the analytical continuation of the potential in its interior or on its surface (this is only another formulation of the definition of σ_0 given above). This presents a striking analogy to the *circle of convergence* in the theory of analytical functions of a complex variable; in fact, real and imaginary parts of such functions form harmonic functions in the plane. Such a circle of convergence separates the regions of convergence and of divergence of complex power series (which are analogues of spatial spherical-harmonic series):

there is convergence everywhere inside this circle and divergence everywhere outside, or vice versa.

It is thus tempting to assume that a similar situation holds in the three-dimensional case: the limit sphere σ_0 is a *sphere of convergence*, separating the regions of convergence (outside σ_0) and of divergence (inside σ_0). This was tacitly assumed as a matter of fact in (Moritz, 1961).

Unfortunately, this is not true, at least not completely, as an elegant counterexample given by Krarup (1969, p.47-49) shows. In fact, there are, in three dimensions, other "surfaces of convergence" besides spheres. For instance, consider, in three dimensions, a harmonic function of two variables x, y only:

$$V = f(x, y) . \quad (6-8)$$

As a harmonic function in two variables, it can be expanded into a Fourier series (cf. Kellogg, 1929, p.353)

$$f(x, y) = \sum_{n=0}^{\infty} \rho^n (a_n \cos n\lambda + b_n \sin n\lambda) \quad (6-9)$$

$$(\rho = \sqrt{x^2 + y^2}, \quad \lambda = \arctan \frac{y}{x}) ,$$

which is the plane equivalent of an expansion into spherical harmonics. For such an expansion, there exists a circle of convergence: convergence inside, divergence (almost everywhere) outside; see below. The unit of length will be chosen such that the circle of convergence is the unit circle

$$x^2 + y^2 = 1 . \quad (6-10)$$

Let us now consider this function of x and y a function V in space; it will clearly be harmonic also as a spatial function. We introduce spherical coordinates r, θ, λ in the usual way by (3-1). Then

$$\rho^n = r^n \sin^n \theta = 2^n \frac{n!}{(2n)!} r^n P_{nn}(\cos \theta) , \quad (6-11)$$

where $P_{nn}(\cos \theta)$ is the sectorial Legendre function; in fact, from (3-7) we get for $m = n$ and $t = \cos \theta$ (there is now $v = 0$):

$$P_{nn}(\cos \theta) = 2^{-n} \sin^n \theta \frac{(2n)!}{n!} \quad (6-12)$$

Substituting (6-11) into (6-9), we obtain

$$V = \sum_{n=0}^{\infty} r^n P_{nn}(\cos\theta) (A_n \cos n\lambda + B_n \sin n\lambda), \quad (6-13)$$

where

$$A_n = 2^n \frac{n!}{(2n)!} a_n, \quad B_n = 2^n \frac{n!}{(2n)!} b_n. \quad (6-14)$$

The "surface of convergence" of this expansion is a cylinder, because (6-10), in space, is the equation of a cylinder.

An inversion in the unit sphere is a transformation

$$x = \frac{x'}{r'^2}, \quad y = \frac{y'}{r'^2}, \quad z = \frac{z'}{r'^2}, \quad rr' = 1, \quad (6-15)$$

where

$$r^2 = x^2 + y^2 + z^2, \quad r'^2 = x'^2 + y'^2 + z'^2. \quad (6-16)$$

It transforms a function $U(x, y, z)$ harmonic in a domain D into a function

$$V(x', y', z') = \frac{1}{r'} U\left(\frac{x'}{r'^2}, \frac{y'}{r'^2}, \frac{z'}{r'^2}\right), \quad (6-17)$$

which is harmonic in the domain D' into which D is carried by the inversion (6-15). This transformation of harmonic functions is called a *Kelvin transformation* (Kellogg, 1929, p.232).

By a Kelvin transformation, the function (6-13) is transformed into the expression

$$V' = \sum_{n=0}^{\infty} \frac{P_{nn}(\cos\theta)}{r^{n+1}} (A_n \cos n\lambda + B_n \sin n\lambda), \quad (6-18)$$

with the same coefficients A_n and B_n . The cylinder (6-10) is transformed into the torus obtained by rotating the circle

$$\left(x - \frac{1}{2}\right)^2 + z^2 = \frac{1}{4}, \quad y = 0 \quad (6-19)$$

around the z -axis (Fig. 6.5). (After transformation, we have again written x, y, z in the place of x', y', z' .) For the series (6-18), this torus is now the convergence surface: the series converges everywhere outside and diverges (almost) everywhere inside this torus.

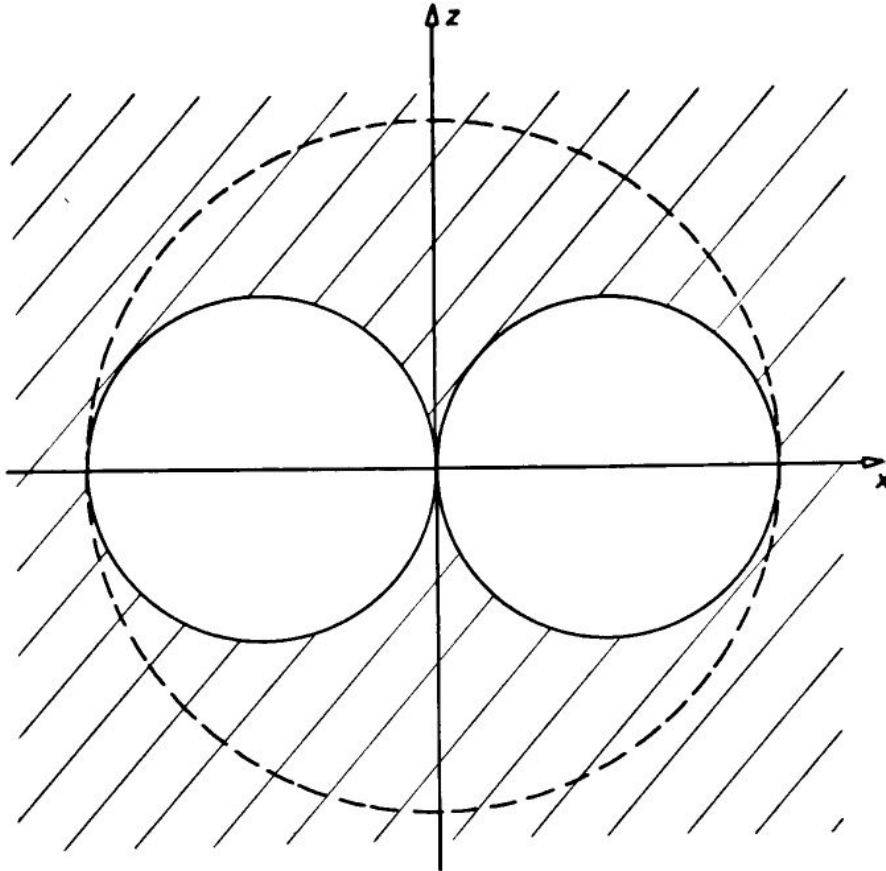


FIGURE 6.5. The torus as a convergence surface.

Thus the purely sectorial expansions (6-13) and (6-14) have a convergence cylinder and a convergence torus, respectively. There are thus special spherical-harmonic series that have convergence surfaces other than spheres.

We still have to show that in two dimensions there is always a circle of convergence: convergence inside, divergence (almost everywhere) outside; we have used this fact in operating with the convergence circle (6-10).

To the harmonic function (6-9) let us associate the complex function

$$F(z) = \sum_{n=0}^{\infty} c_n z^n \quad (6-20)$$

with

$$z = x+iy, \quad c_n = a_n - ib_n, \quad x = \rho \cos \lambda, \quad y = \rho \sin \lambda. \quad (6-21)$$

Then it is easy to show (cf. Kellogg, 1929, p.353) that the function $f(x,y)$ as defined by (6-9) is nothing else than the real part of $F(z)$. It is a basic fact of complex power series that they always have a convergence circle: convergence everywhere inside, divergence everywhere outside. Let the convergence circle of (6-20) be the unit circle (this can be achieved by choosing the unit of length accordingly). Then

$$\limsup \sqrt[n]{|c_n|} = 1 \quad (6-22)$$

(cf. Knopp, 1964, pp.155, 415).

The proof for harmonic functions in the plane is based on the *Cantor-Lebesgue Theorem* (Natanson, 1969, p.334) which, applied to (6-9), states that if this series is to converge on a set of nonzero measure, there must be

$$\rho^n a_n \rightarrow 0, \quad \rho^n b_n \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (6-23)$$

Assume that the series converges for $\rho > 1$ on a set of nonzero measure. Then there must be:

$$\rho^n |c_n| = \rho^n \sqrt{a_n^2 + b_n^2} \rightarrow 0 \quad (6-24)$$

by (6-23). On the other hand, from (6-22) we have if $\rho > 1$:

$$\limsup \sqrt[n]{|c_n|} > \frac{1}{\rho}$$

or

$$\limsup \sqrt[n]{\rho^n |c_n|} > 1, \quad (6-25)$$

which means that $\rho^n |c_n| > 1$ infinitely often, in contradiction to (6-24). This contradiction shows that the Fourier series (6-9) cannot converge outside the unit circle except on a set of measure zero, in other words, for $\rho > 1$ it is divergent almost everywhere.

That the series (6-9) is everywhere convergent inside the unit circle is immediately seen by writing it as

$$f(x,y) = \sum_{n=0}^{\infty} \rho^n |c_n| \cos(n\lambda + \beta_n) \quad (6-26)$$

and noting that $\cos(n\lambda + \beta_n) \leq 1$. Therefore, this series is majorized by

$$\sum_{n=0}^{\infty} |c_n| \rho^n \quad (6-27)$$

which has the same radius of convergence as (6-20), namely 1.

Note that, whereas in the case of power series (6-20) there is divergence *everywhere* outside the convergence circle, in the case of Fourier series (6-9) we can assert divergence only *almost everywhere* outside the convergence circle; there may be convergence on a set of measure zero even outside the circle of convergence. For instance, if in (6-9) all $a_n = 0$, then the series converges everywhere on the x-axis to the value $f(x,y) = 0$ because $\sin n\lambda = 0$ for $\lambda = 0$; this is obviously true even if $\rho > 1$. We thus have convergence on an infinite straight line, which constitutes a set of measure zero in the plane (for the concept of measure, cf. (Kolmogorov and Fomin, 1970, sec.25)).

Let us now return to spherical harmonics in space. We have seen that there may be surfaces of convergence other than a sphere, such as a cylinder or a torus. However, it has been conjectured (Krarup, 1969, p.49) that these cases are exceptional and that, as a rule, the surface of convergence is a sphere. In particular, for *zonal harmonics* (Legendre polynomials), there should be a convergence sphere.

The case of a series of zonal harmonics has been investigated by Ecker (1970a,b), who has tried to find a theorem of Cantor-Lebesgue type. He has adduced arguments which, although not mathematically rigorous, make the existence of a convergence sphere for zonal harmonics highly probable. Ecker (1970b) has also studied convergence surfaces of form $r = C \sin^l \theta$ of which sphere ($l=0$), torus ($l=1$), and cylinder ($l=-1$) are special cases.

We finally give an example of a zonal harmonic series for which there is a sphere of convergence (Baeschlin, 1948, p.430) and which we shall need later on. Consider a mass point A situated on the z-axis at a distance r' from the origin O (Fig. 6.6). By (3-32) the potential due to this point (of mass m) is

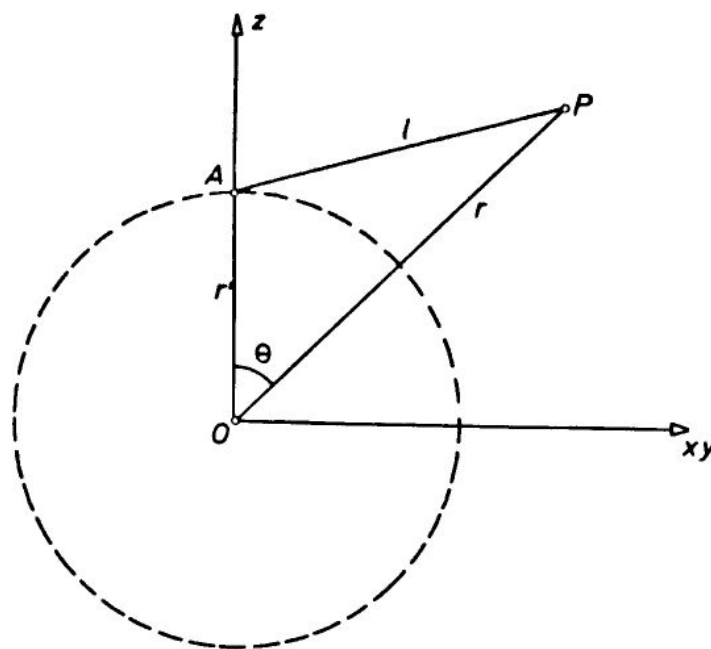


FIGURE 6.6. The potential of a point mass at A.

$$v = \frac{Gm}{l} = Gm \sum_{n=0}^{\infty} \frac{r^n}{r^{n+1}} P_n(\cos \theta), \quad (6-28)$$

which represents a spherical harmonic expansion consisting only of zonal harmonics. Putting

$$\frac{r^n}{r} = z \quad (6-29)$$

we see that, apart from a factor, this series equals

$$\sum_{n=0}^{\infty} z^n P_n(\cos \theta) = \frac{1}{\sqrt{1-2z \cos \theta + z^2}}. \quad (6-30)$$

Considering, for a moment, z as a complex variable, the series (6-30) has a convergence circle $|z| = \text{const.}$ The only singularities of the function of z represented by the right-hand side of (6-30) are the values

$$z_1, z_2 = e^{\pm i\theta}, \quad (6-31)$$

for which the denominator is zero. Both singularities z_1 and z_2 lie on the unit circle

$$|z| = 1, \quad (6-32)$$

which is, therefore, the convergence circle for the complex power series (6-30).

By a well-known relation between complex and real power series, 1 is also the radius of convergence if z is real. Therefore, by (6-29), the series (6-28) is convergent if $r > r'$ and divergent if $r < r'$, whatever the value of θ is. Hence, the sphere $r' = r$ through A is a true *convergence sphere* separating the region of convergence and divergence.

We thus see that the problem of convergence of the spherical-harmonic expansion of the external gravitational potential at the earth's surface remains open, for two reasons:

1. The behavior of the analytical continuation of the external potential into the earth's interior is not known.
2. We do not know the precise form of the convergence surface for the spherical-harmonic series of the earth's potential.

In the next section we shall take up the convergence problem from a completely different angle.

7. CONVERGENCE OF SPHERICAL HARMONICS II

In this section we shall arrive at the surprising conclusion that the convergence problem for spherical harmonics is physically and practically almost meaningless.

In sec. 6 we have seen that, in the case of the equipotential ellipsoid, the corresponding spherical harmonic expansion does converge at the ellipsoid (and still further down). Since the earth is very nearly an ellipsoid, it would be tempting to conclude "by analogy" that the actual spherical-harmonic series also converges, at least down to the earth's surface.

Such a conclusion would be completely unwarranted, however. The reason is that analytical continuation is an extremely unstable property. A small grain of sand is sufficient to completely change the situation.

This can be taken literally (Fig. 7.1). Assume the earth to be an exact level ellipsoid, for which the series converges throughout its surface E .

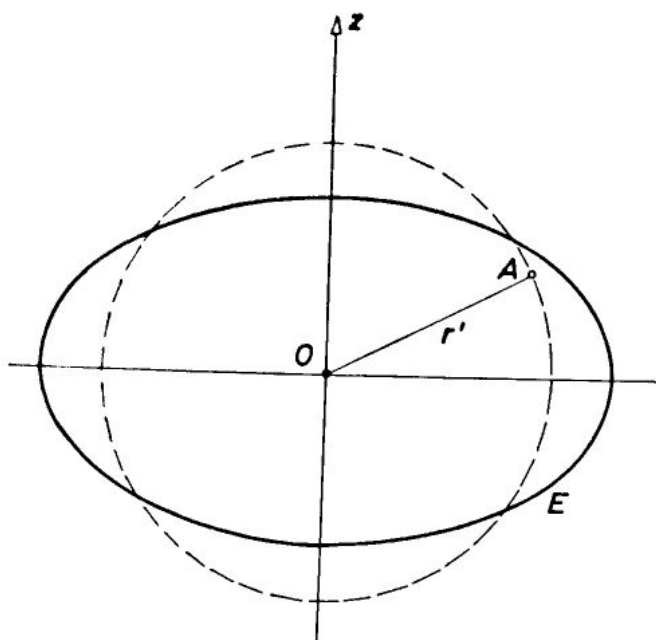


FIGURE 7.1. A grain of sand changes convergence.

Now take a small "grain of sand" of the form of a mass point, or of a small homogeneous sphere for which the external gravitational potential is identical to that of a mass point. Bury this grain, of mass m , at a point A which is situated at a small distance below the surface. The potential v of this grain is now given by the series (6-28), for which the sphere $r = r'$ is the convergence sphere. Since all singularities of the external ellipsoidal potential V and its analytical continuation lie inside this sphere (p.54), the sphere $r = r'$ is also the convergence sphere for the "perturbed" potential $V + v$. (The fact that the expansions refer to different axes, V to Oz and v to OA , is irrelevant in view of the invariance of spherical harmonics with respect to rotation.)

Since the ellipsoidal surface lies partly inside and partly outside the sphere of convergence, there is now a part of the ellipsoid on which the spherical-harmonic expansion of the "perturbed" potential $V + v$ diverges.

It is clear that by selecting the mass m very small, the perturbation v of the potential can be made, outside and on the ellipsoid, as small as desired. Thus, an arbitrarily small change of the external potential may change convergence into divergence. Convergence is a very unstable property indeed.

It is now very remarkable that the instability argument can also be reversed. This is possible through an application of Runge's theorem (Krarp, 1969, p.54).

Runge's theorem will be discussed in the following section. Applied to our present problem, it says that the functions ψ , regular and harmonic outside and on a sphere σ_2 completely inside the earth, are dense within the set of functions ϕ , regular and harmonic outside the earth's surface S ; cf. Fig. 6.3. More precisely, every function ϕ can be approximated by a function ψ such that the relation

$$|\phi - \psi| < \epsilon, \quad (7-1)$$

for arbitrarily small positive ϵ , holds everywhere outside and on a closed surface S_1 which surrounds the earth's surface S and is arbitrarily close to it. If S is sufficiently smooth, S_1 may even be identified with S (theorem of Keldysh-Lavrentiev).

In our case, ϕ is the earth's external gravitational potential, which is harmonic and regular outside the earth's surface S ; inside S , it may have singularities. Nevertheless, we can always without committing an appreciable error (we may take ϵ much smaller than any possible empirical uncertainty), approximate the external potential ϕ by a harmonic function ψ that is regular down to some given sphere completely inside the earth. For ψ , the spherical-harmonic expansion will be convergent at the earth's surface (p.55), even if the corresponding expansion of the original potential ϕ does not converge at this surface. Thus, by an arbitrarily small change of the external potential it is possible to change divergence into convergence.

Together with the fact, shown above, that convergence can be as easily changed into divergence, this shows that convergence (or divergence) at the earth's surface is indeed such an instable property as to be completely meaningless from a physical point of view.

In fact, even apart from uncertainties of empirical determination, the physical definition of the potential can never be completely sharp, in view of the atomic structure of matter and the corresponding quantum-mechanical uncertainties.

A simple example will help to clarify the situation. Let us pose the question whether a certain distance defined in nature, say the distance between two trigonometric stations, represents a rational or an irrational number. Clearly, the question is meaningless: distance is physically definable only within a certain, very small, uncertainty (again because of the atomic structure of matter), and within this range of uncertainty there lie infinitely many rational and infinitely many irrational numbers, all of them perfectly respectable and equally suited candidates for the office of numerically representing the distance under consideration.

Thus it is always possible to regard an empirical quantity as a rational number; even apart from the limited measuring accuracy, which further makes the choice of a rational number expressed by a finite decimal expression completely natural.

The mathematical reason behind this is, of course, the fact, that every irrational number can be approximated with arbitrary accuracy by a rational number (and vice versa); the rational numbers are *dense* within the set of real numbers (p.42).

The question of convergence or divergence, at the earth's surface, of the spherical-harmonic expansion of the external potential is completely analogous to the question whether a certain observable quantity has a rational or irrational numerical value. In fact, let us for brevity introduce the name "convergent potential" for an external potential whose spherical-harmonic expansion is convergent on and outside the earth's surface. Then the theorem by Runge-Krarup states that the set of "convergent potentials" is *dense* within the set of all possible external potentials, just as the rationals are dense within the set of real numbers.

Thus our question of convergence or divergence is as meaningless as the question of rationality or irrationality of a physically defined numerical value.

As a practical consequence we recognize that it is always possible to consider the earth's external potential as a "convergent potential", in the same way as it is always possible to consider the result of a certain measurement as a rational number.

Uniform approximation by spherical harmonics. Although the external potential ϕ cannot, in general, be expanded into a spherical harmonic series which converges at the earth's surface, it can be approximated, uniformly on and outside this surface, by a finite linear combination of spherical harmonics ψ_n . In fact, such linear combinations are dense within the set of "convergent potentials" ψ : suitable linear combinations are simply obtained by truncating the corresponding spherical-harmonic series, which is convergent in this case. We can thus, for each ψ , find a ψ_n for which

$$|\psi - \psi_n| < \frac{\epsilon}{2}. \quad (7-2)$$

Since the functions ψ are dense in the space of functions ϕ , we can find a ψ such that

$$|\phi - \psi| < \frac{\epsilon}{2}. \quad (7-3)$$

Combining the two inequalities we get

$$|\phi - \psi_n| \leq |\phi - \psi| + |\psi - \psi_n| < \epsilon, \quad (7-4)$$

which was to be shown.

8. RUNGE'S THEOREM

In the convergence study of sec. 7, in the solution of Molodensky's problem (sec. 45), and, in particular, throughout least-squares collocation (cf. p. 98), we wish to approximate the earth's external gravitational potential by a harmonic function that is regular down to sea level or even down to a sphere that lies completely inside the earth. The possibility of such an approximation is guaranteed by *Runge's theorem*:

Let K be a compact set and Γ and Ω open sets in R^3 , such that their boundaries are homeomorphic to a sphere and such that $K \subset \Gamma$ and $\Gamma \subset \Omega$. If the function ϕ is harmonic in Γ and if $\epsilon > 0$ is arbitrarily small, then there exists a function ψ , harmonic in Ω , such that

$$|\phi - \psi| < \epsilon$$

(8-1)

uniformly on K .

The regions K , Γ , and Ω are illustrated by Fig. 8.1. The set K is compact (i.e., closed and bounded) and the sets Γ and Ω are open; $\bar{\Gamma}$ is the closure of Γ : it consists of the open set Γ together with its boundary. "A surface is homeomorphic to a sphere" means that it is a closed boundary. "A surface is homeomorphic to a sphere" means that it is a closed surface which can be continuously deformed into a sphere (such as the ellipsoid or the earth's surface). The condition $K \subset \Gamma$ says that K is completely contained in the open set Γ , and similarly with $\bar{\Gamma} \subset \Omega$. These two conditions prevent the bounding surfaces from touching or intersecting each other. Note that Ω is the whole interior of the outermost surface and that Γ is the whole interior of the middle surface in Fig. 8.1.

When we say that a function is harmonic in a region Γ , then we mean that it also is regular in this region. Thus Runge's theorem says that a function regular and harmonic only in an open region Γ (it may have singularities already on the boundary of Γ !), can be arbitrarily well approximated by a function regular and harmonic in a larger region Ω .

As it stands, and with the sets K , Γ , and Ω situated as in Fig. 8.1, this theorem does not seem to have much geodetic relevance. However, Runge's theorem holds also when the regions K , Γ , and Ω are the exteriors of

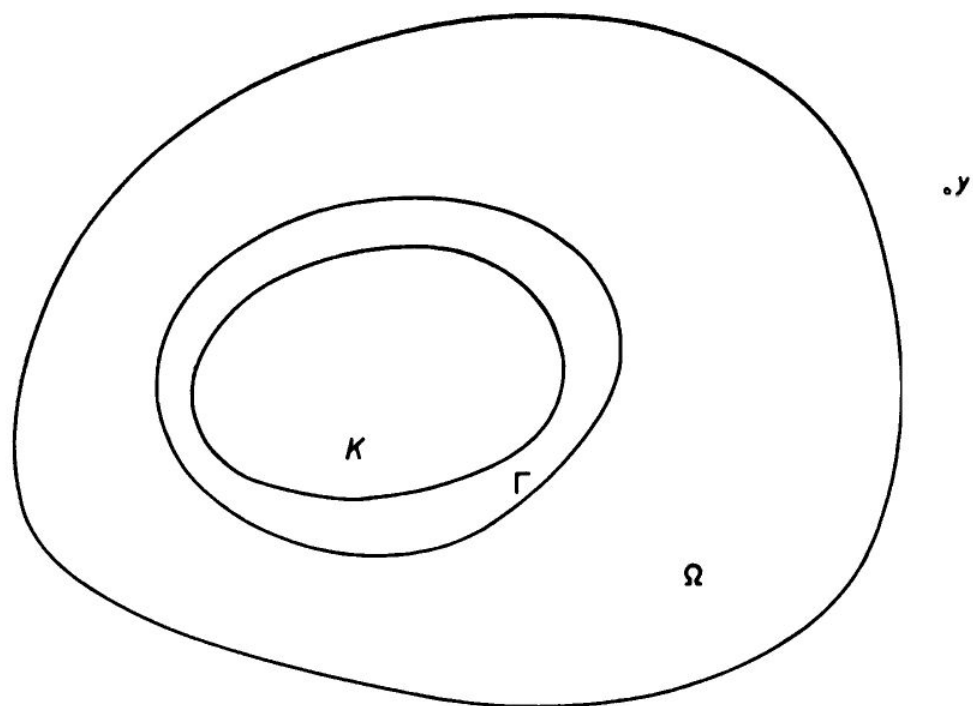


FIGURE 8.1. The compact set K and the open sets Γ and Ω .

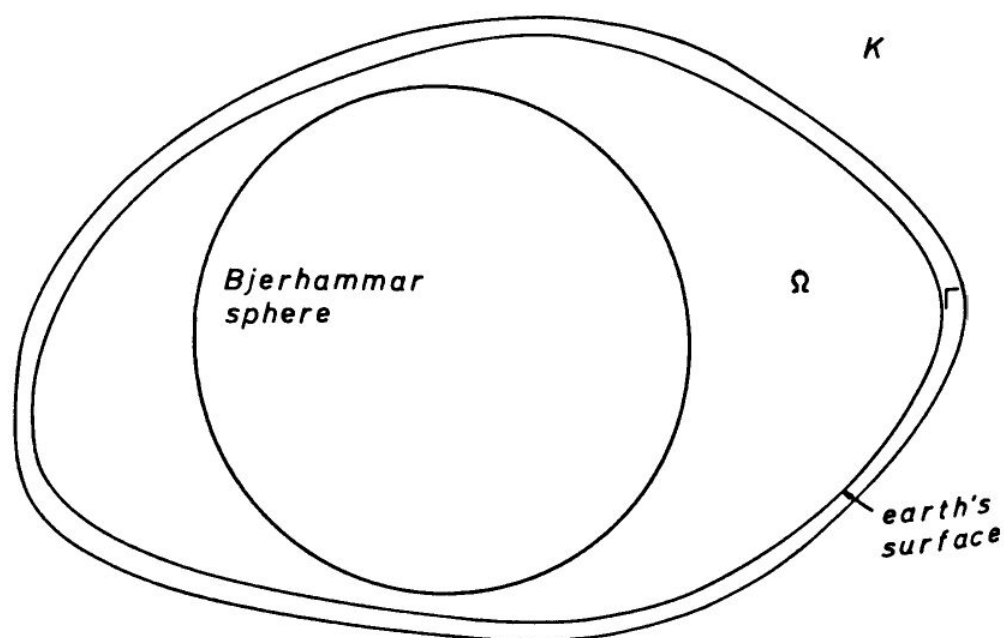


FIGURE 8.2. The geodetically relevant case.

their respective boundary surfaces (Fig. 8.2). In this geodetically relevant case we may identify r with the exterior of the physical earth's surface, Ω with the exterior of some sphere that is completely inside the earth (also called Bjerhammar sphere), and K with the exterior of some surface, completely enclosing the earth (more strictly, K is the exterior plus the bounding surface since K is a closed set; it is no longer compact since it is unbounded).

Denoting the boundaries of K , r , and Ω by ∂K , ∂r , and $\partial \Omega$, respectively, then ∂r is the earth's surface, $\partial \Omega$ is the surface of the Bjerhammar sphere, and ∂K is a surface completely surrounding the earth's surface and arbitrarily close to it.

We then have the geodetic version of Runge's theorem, which we shall call the *Runge-Krarup-theorem* (Krarup, 1969, p.54; Krarup, 1975):

Any harmonic function ϕ , regular outside the earth's surface, may be uniformly approximated by harmonic functions ψ regular outside an arbitrarily given sphere inside the earth, in the sense that for any given $\epsilon > 0$, the relation

$$|\phi - \psi| < \epsilon$$

holds everywhere outside and on any closed surface completely surrounding the earth's surface.

The number ϵ may be arbitrarily small, and the surrounding surface ∂K may be arbitrarily close to the earth's surface ∂r ; the two surfaces may, e.g., nowhere differ by more than 0.1 mm. This is clearly sufficient for any practical application, but if the earth's surface is sufficiently regular (e.g., continuously differentiable), the surface ∂K may even be taken to coincide with ∂r . We then get the *Keldysh-Lavrentiev theorem* (Bjerhammar, 1975):

Any function ϕ , harmonic outside the earth's surface and continuous outside and on it, may be uniformly approximated by harmonic functions ψ regular outside an arbitrarily given sphere inside the earth, in the sense that for any given $\epsilon > 0$, the relation

$$|\phi - \psi| < \epsilon$$

holds everywhere outside and on the earth's surface.

Note that if the condition $|\phi - \psi| < \epsilon$ is satisfied on the boundary, it is automatically satisfied in the whole exterior, because a harmonic function attains its maximum and minimum values only at the boundary (maximum principle, cf. (Kellogg, 1929), p.223).

Since ϵ can be chosen arbitrarily, we may consider a sequence

$$\epsilon_1 > \epsilon_2 > \epsilon_3 \dots \rightarrow 0 \quad (8-2)$$

and find a function ψ_k for each ϵ_k . In this way, we obtain a sequence of harmonic functions $\psi_1, \psi_2, \psi_3, \dots$, regular in the whole space outside the given sphere $\partial\Omega$ and converging uniformly to the given function ϕ on and outside ∂K (or $\partial\Gamma$, respectively).

Uniform convergence on K corresponds to the uniform norm (5-2). We may thus work with the space C of functions continuous on the closed set K ; the norm is defined by (5-2), the maximum being taken over K .

Using this terminology, we may regard Runge's theorem in its various forms as a *denseness theorem*. In the Keldysh-Lavrentiev version we may thus say: The harmonic functions ψ admitting a regular continuation down to the Bjerhammar sphere, are dense in the subspace of C formed by the functions harmonic outside the earth's surface and continuous outside and on it; the space C itself is formed by the functions continuous on and outside the earth's surface (and behaving suitably at infinity).

Proof of Runge's theorem. This proof is conceptually rather demanding; it may be omitted by readers not interested in mathematical details. Our present proof is modeled after the proof of Runge's theorem for functions of a complex variable in (Hörmander, 1966, pp.6-8) and is essentially a detailed version of the proof in (Bjerhammar, 1975). We shall first prove the "interior" version, given at the very beginning of the section, for the situation illustrated in Fig. 8.1.

The basic tool is the *Hahn-Banach theorem* fundamental in functional analysis; we shall use it in the version of Theorem 4 in chapter IV, § 2 of (Kantorovich and Akilov, 1964):

Let X be a normed space and E an arbitrary subset of X . An element $f_0 \in X$ belongs to the linear closure of the set E if and only if

$$L(f_0) = 0 \quad (8-3)$$

for all functionals vanishing on E .

The linear closure Y of E consists of all elements of X which can be represented as finite or infinite linear combinations of E . Thus, Y is the linear subspace spanned by the elements of E , that is, the smallest linear subspace of X which contains all elements of E .

Geometrically, this theorem can be interpreted as follows. Eq. (8-3), for a definite functional L , defines a plane (or hyperplane) passing through the origin in the space X . (To see this, take the special case of

\mathbb{R}^n in which the linear functional is given by (4-52).) Each functional (8-3) vanishing on E defines such a plane, and the subspace Y is then formed by the intersection of all these planes. If x_0 is to belong to the subspace Y , then evidently (8-3) must be satisfied.

This geometrical consideration does not replace a proof, which the reader may find in the book just quoted, but will provide an intuitive understanding of this basic theorem.

In the present case, X is the space C of continuous functions on K , the subset E is the set of functions $H(\Omega)$ harmonic in Ω (it is a subset of C because each function from $H(\Omega)$ is clearly continuous on K , thus belongs to C), and f_0 is a function harmonic in the smaller region Γ . In the terminology of the formulation of the Runge theorem this means that

$$\phi = f_0, \quad \psi \in H(\Omega) = E. \quad (8-4)$$

We have to prove that ϕ can be approximated arbitrarily well by elements $\psi \in H(\Omega)$, which means that $\phi = f_0$ belongs to the linear closure of E (ϕ does not belong to $H(\Omega)$ since it is harmonic only in Γ and not, in general, in the whole domain Ω !).

By (5-17), the condition (8-3) can be written

$$L(f_0) = \iiint_K \phi(x) dv(x) = 0 \quad (8-5)$$

$\nu(x)$ denoting a signed measure. Geometrically we may say that the measure $\nu(x)$ is *orthogonal* to the function $\phi(x)$; this concept of orthogonality is a generalization of orthogonality of two vectors h and x in \mathbb{R}^n which means that the inner product (4-53) vanishes.

What we, therefore, have to prove for Runge's theorem is that every measure μ orthogonal to $H(\Omega)$ is also orthogonal to every function ϕ harmonic in Γ . This means that from

$$\iiint_K \psi dv = 0 \quad \text{for every } \psi \in H(\Omega) \quad (8-6)$$

it must follow that also

$$\iiint_K \phi dv = 0. \quad (8-7)$$

Assume that (8-6) holds and consider a point y outside $\bar{\Omega}$ (Fig. 8.1).¹ Then, denoting the distance between y and a point $x \in K$ by $|x - y|$, the function

$$\psi(x) = \frac{1}{|x - y|} \quad (8-8)$$

for fixed y and variable x is harmonic in Ω , i.e., belongs to $H(\Omega)$. From (8-6) it follows that

$$V(y) = \iiint_K \frac{dv(x)}{|x - y|} \quad (8-9)$$

is zero, as a function of y , everywhere outside $\bar{\Omega}$.

The physical interpretation of the function $V(y)$ is, of course, that of the potential generated by the (positive or negative) masses $v(x)$, the gravitational constant G being set equal to 1. This potential is a harmonic function outside the generating masses which are concentrated on K , that is, it is harmonic outside K and vanishes identically outside Ω . Since the region $\Omega - K$ (the part of Ω different from K) is connected, there exists a unique analytical continuation of $V(y)$ from the outside of Ω into $\Omega - K$, which must be identically zero since $V(y)$ is identically zero outside $\bar{\Omega}$.

Hence

$$V(y) \equiv 0 \text{ outside } K. \quad (8-10)$$

Inside K , $V(y) \neq 0$ since V is not harmonic there. The function $V(y)$ is the potential of a mass distribution generating a zero outer potential. Such distributions are well known; an example is furnished by two homogeneous concentric spherical shells of the same total charge, one being positive and the other negative.

Consider now the function ϕ harmonic in τ , for which we wish to prove (8-7), and an auxiliary function $h(x)$, $x \in R^3$, such that

$$h(x) \equiv 1 \text{ on } K, \quad h(x) \equiv 0 \text{ outside } K_1, \quad (8-11)$$

K_1 being a compact set contained in τ but containing K in its interior;

¹This means that y lies neither in Ω nor at its boundary.

in the region $K_1 - K$ the function $h(x)$ is interpolated so smoothly that $h(x)$ is infinitely differentiable in the whole space R^3 . Such functions (called infinitely differentiable functions of compact support) exist and play a great role in modern mathematics; cf. (Wladimirow, 1972, pp.68-70).

Then the product of these two functions,

$$U(x) = h(x)\phi(x) , \quad (8-12)$$

may be extended naturally into the whole space R^3 and has then the following properties resulting from (8-11) and from the harmonicity of ϕ :

$$\begin{aligned} U(x) &= \phi(x) , \quad \Delta U = \Delta \phi = 0 \quad \text{on } K , \\ U(x) &\equiv 0 \quad \text{outside } \Gamma ; \end{aligned} \quad (8-13)$$

$U(x)$ is infinitely differentiable in the whole space R^3 .

According to (Kellogg, 1929, p.219), any function twice continuously differentiable in $\bar{\Gamma}$ can be represented in the form

$$4\pi U(x) = -\iiint_{\Gamma} \frac{\Delta U}{l} d\Gamma + \iint_{\partial\Gamma} \frac{\partial U}{\partial n} \frac{1}{l} dS - \iint_{\partial\Gamma} U \frac{\partial}{\partial n} \left(\frac{1}{l} \right) dS . \quad (8-14)$$

Here l is the distance between the point x and the volume element $d\Gamma$ or the surface element dS , respectively; $\partial/\partial n$ denotes the derivative with respect to the outer normal to the surface $\partial\Gamma$. This boundary surface must be sufficiently regular, which can always be assumed (if necessary, this condition can be fulfilled by a slight deformation of $\partial\Gamma$ "inward"). The physical interpretation of (8-14) is clearly a representation of $U(x)$ as the sum of three potentials: of a volume distribution, a surface layer, and a double layer.

In the present case (8-13), the surface layer and the double layer are zero because $U(x)$ is zero, together with all derivatives, at the surface $\partial\Gamma$, and there remains the volume distribution:

$$U(x) = -\frac{1}{4\pi} \iiint_{\Gamma} \frac{\Delta U}{l} d\Gamma . \quad (8-15)$$

Let us now denote by z the point at which $d\Gamma$ is situated, so that $l = |x - z|$. Then (8-15) takes the form

$$U(x) = -\frac{1}{4\pi} \iiint_{\Gamma} \frac{\Delta U(z)}{|x-z|} d\Gamma(z). \quad (8-16)$$

Let us now consider (8-7). Using (8-13) and (8-16) we have

$$\iiint_K \phi(x) dv(x) = \iiint_K U(x) dv(x) = -\frac{1}{4\pi} \iiint_K \iiint_{\Gamma} \frac{\Delta U(z)}{|x-z|} d\Gamma(z) dv(x).$$

The order of the two integrations can be inverted (Fubini's theorem):

$$\iiint_K \phi(x) dv(x) = -\frac{1}{4\pi} \iiint_{\Gamma} \left[\iiint_K \frac{dv(x)}{|x-z|} \right] \Delta U(z) d\Gamma(z).$$

The integral between the square brackets equals $V(z)$, by (8-9), so that

$$\begin{aligned} \iiint_K \phi(x) dv(x) &= -\frac{1}{4\pi} \iiint_{\Gamma} V(z) \Delta U(z) d\Gamma(z) \\ &= -\frac{1}{4\pi} \iiint_K V \Delta U d\Gamma - \frac{1}{4\pi} \iiint_{\Gamma-K} V \Delta U d\Gamma. \end{aligned}$$

The first integral is zero because $\Delta U = 0$ on K , by (8-13), and the second integral is zero because $V = 0$ outside K by (8-10).

Thus (8-7) is, indeed, a consequence of (8-6), which completes the proof of Runge's theorem.

The Keldysh-Lavrentiev theorem follows from the Runge theorem using Theorem 5.18 in (Landkof, 1972, p.341), which states that, for very general compact regions K , the linear manifold \underline{H}_K of functions, harmonic in some neighborhood of K , is dense in the space H_K of functions continuous on K (i.e., including the boundary) and harmonic in the interior of K ; the norm is always the uniform norm over K . Since Runge's theorem says that $H(\Omega)$ is dense in \underline{H}_K , it follows that $H(\Omega)$ is dense in H_K . Hence, every function $\in H_K$ can be uniformly approximated by functions $\in H(\Omega)$, which is the Keldysh-Lavrentiev theorem.

A direct proof of the latter theorem is again found in (Bjerhammar, 1975).

So far we have restricted ourselves to interior regions. The corresponding theorems for the geodetically relevant case of exterior regions follow simply by a Kelvin transformation (Kellogg, 1929, p.232).

Elementary proofs of Runge's theorem may be found in J.L. Walsh, Bull. Amer. Math. Soc., 35 (1929), pp. 499-544, and in Ph. Frank and R. von Mises, Die Differential- und Integralgleichungen der Mechanik und Physik, vol.1, (1930; reprint by Dover Publ., New York, 1961), p. 760.

PART B

LEAST-SQUARES COLLOCATION: ELEMENTARY APPROACH

Least-squares collocation is a method for determining the anomalous gravitational field by a combination of geodetic measurements of different kinds. In this Part B we shall present the subject in an elementary way, starting from least-squares prediction familiar from gravity interpolation and emphasizing relations to least-squares adjustment. Although we shall approach the topic from a statistical point of view, the analytical structure, which our method shares with collocation methods in approximation theory and which is essential for physical geodesy, is also stressed (sec.12).

Our treatment will be inductive, progressing from the simplest situation to more complex cases by successive generalizations. After an elementary presentation of the prediction problem we consider the "pure" case of collocation without random errors and systematic effects (sec.11). Random errors are introduced in sec. 14, and sec. 16 treats a general model which is a synthesis between least-squares prediction, collocation in the sense of approximation theory, and least-squares adjustment. Various applications to problems of physical geodesy are discussed.

Least-squares collocation, in the same way as certain adjustment problems, may lead to very large systems of linear equations. Therefore, step-wise techniques can become important; they are presented in sections 19 and 20.

A fundamental role is played by the covariance function of the anomalous gravitational potential. Its definition and basic properties are given in sec. 10; a more detailed discussion from a mathematical and numerical point of view will be found in sections 22 and 23.

The treatment in Part B is elementary in the sense that functional analysis is almost completely bypassed and only linear algebra is used; it is thus oriented towards application.

9. LEAST-SQUARES PREDICTION

Assume two sets of random quantities: the set of "measurements" l_1, l_2, \dots, l_q forming the q -vector

$$l = [l_1 \ l_2 \ \dots \ l_q]^T \quad (9-1)$$

and the set of "signals" s_1, s_2, \dots, s_m forming the m -vector

$$s = [s_1 \ s_2 \ \dots \ s_m]^T ; \quad (9-2)$$

the superscript T denotes transposition, as usual, so that l and s are column vectors, and commas between vector components are omitted.

It is assumed that each of these quantities has an expected value equal to zero:

$$E\{l\} = 0 , \quad E\{s\} = 0 , \quad (9-3)$$

the expectation $E\{\cdot\}$ being the average or mean value in the sense of probability theory; cf. (Liebelt, 1967, p.85). Quantities having mean value zero, such as (9-3), are called *centered*.

We also consider the *covariance matrices*

$$C_{ll} = \text{cov}(l, l) , \quad (9-4)$$

$$C_{sl} = \text{cov}(s, l) , \quad (9-5)$$

$$C_{ss} = \text{cov}(s, s) . \quad (9-6)$$

C_{ll} and C_{ss} are the autocovariance matrices of the vectors l and s , respectively, and C_{sl} is the cross-covariance matrix between l and s . The elements of the $q \times q$ matrix C_{ll} are the average products

$$E\{l_i l_j\} , \quad i, j = 1, 2, \dots, q ; \quad (9-7)$$

the elements of the $m \times q$ matrix C_{sl} are

$$E\{s_k l_i\} , \quad k = 1, 2, \dots, m ; \quad (9-8)$$

and the elements of the $m \times m$ matrix C_{ss} are

$$E\{s_k s_h\}, \quad k, h = 1, 2, \dots, m. \quad (9-9)$$

This is true because our random quantities are centered; cf. (*ibid.*, p.95). In vector notation we may thus write:

$$C_{ll} = E\{ll^T\}, \quad (9-10)$$

$$C_{sl} = E\{sl^T\}, \quad (9-11)$$

$$C_{ss} = E\{ss^T\}. \quad (9-12)$$

It is supposed that these covariance matrices, and all other matrices occurring in this book, have full rank; an $m \times n$ matrix A is said to have full rank if $\text{rank}(A) = m$ or n whichever is smaller.

The measurement vector l is assumed to be known, the signal vector s is unknown. What is the best estimate for s on the basis of the data l ? The connection between s and l is given, not through a functional relation, but only in terms of the covariance matrices (9-10) to (9-12).

A linear estimate for the vector s has the form

$$\hat{s} = Hl, \quad (9-13)$$

where H is some $m \times q$ matrix; that is, each component of the vector s is approximated by a linear combination of the data l .

The error vector ϵ is given by

$$\epsilon = \hat{s} - s; \quad (9-14)$$

its covariance matrix

$$C_{\epsilon\epsilon} = \text{cov}(\epsilon, \epsilon) = E\{\epsilon\epsilon^T\} = E\{(\hat{s} - s)(\hat{s} - s)^T\} \quad (9-15)$$

is called *error covariance matrix*. The diagonal terms of this matrix are the *error variances* σ_k^2 of the estimated signals \hat{s}_k , which are the components of the vector \hat{s} :

$$\sigma_k^2 = E\{\epsilon_k^2\} = E\{(\hat{s}_k - s_k)^2\}. \quad (9-16)$$

As the best linear estimate of s in terms of l we define, as usual in statistical estimation, the *linear minimum variance unbiased estimate*. Forming the average of (9-13) we get in view of (9-3)

$$E\{\xi\} = HE(l) = 0 = E\{s\} ; \quad (9-17)$$

this is the natural condition for unbiasedness in the present case. Thus (9-13) can be considered as unbiased for any matrix H .

Let us now try to determine H so that the error variances (9-16) are minimum. First we find the error covariance matrix (9-15) for an arbitrary matrix H .

By (9-13) and (9-14) we get

$$\begin{aligned} \epsilon\epsilon^T &= (Hl-s)(Hl-s)^T = (Hl-s)(l^TH^T-s^T) \\ &= Hll^TH^T - sl^TH^T - Hls^T + ss^T . \end{aligned} \quad (9-18)$$

The average of this expression then gives the error covariance matrix (9-15). We obtain

$$E\{\epsilon\epsilon^T\} = HE\{ll^T\}H^T - E\{sl^T\}H^T - HE\{ls^T\} + E\{ss^T\}$$

or, by (9-10), (9-11), (9-12), and (9-15),

$$C_{\epsilon\epsilon} = HC_{11}H^T - C_{s1}H^T - HC_{1s} + C_{ss} , \quad (9-19)$$

in analogy to (9-11) putting

$$C_{1s} = C_{s1}^T = E\{ls^T\} . \quad (9-20)$$

Eq.(9-19) is equivalent to

$$C_{\epsilon\epsilon} = C_{ss} - C_{s1}C_{11}^{-1}C_{1s} + (H-C_{s1}C_{11}^{-1})C_{11}(H-C_{s1}C_{11}^{-1})^T . \quad (9-21)$$

This is readily verified by performing the multiplications and rearranging, using

$$C_{11}^{-1}C_{11} = I \quad (\text{unit matrix}).$$

The inverse C_{11}^{-1} exists since we have supposed all covariance matrices to have full rank.

The matrix (9-21) is the sum of two matrices:

$$A = C_{ss} - C_{s1}C_{11}^{-1}C_{1s} \quad (9-22)$$

and

$$B = (H - C_{s1}C_{11}^{-1})C_{11}(H - C_{s1}C_{11}^{-1})^T \quad (9-23)$$

The matrix A does not depend on H ; it is thus the same for all possible linear estimates (9-13).

The matrix B can be made zero by putting

$$H = C_{s1}C_{11}^{-1}; \quad (9-24)$$

if this equation is not satisfied, then the diagonal terms of B will always be positive. In fact, denote the k -th row of the matrix $H - C_{s1}C_{11}^{-1}$ by γ , which is a row vector of m components. Then the k -th diagonal term of the matrix (9-23) is given by

$$\gamma C_{11} \gamma^T \quad (9-25)$$

Now it is well known that all regular covariance matrices are *positive definite*. By definition, a $r \times r$ matrix M is positive definite if

$$xMx^T \geq 0 \quad (9-26)$$

for an arbitrary row vector x of r components, the equality sign holding only if $x = 0$. Therefore, the quantity (9-25) must be nonnegative:

$$\gamma C_{11} \gamma^T \geq 0; \quad (9-27)$$

the equality sign holds only if $\gamma = 0$, that is, if (9-24) holds. Thus, in view of

$$C_{\epsilon\epsilon} = A + B,$$

the diagonal terms of $C_{\epsilon\epsilon}$, which form the error variances to be minimized, are always larger than the diagonal terms of A , unless $B = 0$.

The minimum variance estimate is thus given for $B = 0$, so that H is expressed by (9-24). The substitution into (9-13) gives

$$\hat{s} = C_{s1} C_{11}^{-1} l, \quad (9-28)$$

which thus provides the best (unbiased minimum variance) linear estimate of the signal vector s in terms of the data vector l .

Eq.(9-28) will be called the formula of *least-squares prediction* since it is a precise analogue of the Kolmogorov-Wiener prediction formula well known from the theory of stochastic processes, cf. (Grafarend, 1975).

For the optimal estimation (9-28) we have $B = 0$, so that (9-21) reduces to

$$E_{ss} = C_{ss} - C_{s1} C_{11}^{-1} C_{1s}, \quad (9-29)$$

if we write

$$C_{\epsilon\epsilon} = E_{ss} \quad (9-30)$$

to denote the error covariance matrix for the prediction of the signal s (E_{ss} should not be confused with the expectation E !).

Gravity prediction. Least-squares interpolation and extrapolation of gravity anomalies (Heiskanen and Moritz, 1967, sec.7-6) may be considered as an application of the present prediction method. In fact, let

$$l = [\Delta g_1 \ \Delta g_2 \ \dots \ \Delta g_q]^T \quad (9-31)$$

be the known (errorless) gravity anomalies at q observation points P_i , and let

$$s = \Delta g_p \quad (9-32)$$

be the gravity anomaly at an interpolation point P (there is $m = 1$). Then (9-28) becomes

$$\Delta g_P = \begin{bmatrix} C_{P1} & C_{P2} & \dots & C_{Pq} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1q} \\ C_{21} & C_{22} & \dots & C_{2q} \\ \vdots & \vdots & & \vdots \\ C_{q1} & C_{q2} & \dots & C_{qq} \end{bmatrix}^{-1} \begin{bmatrix} \Delta g_1 \\ \Delta g_2 \\ \vdots \\ \Delta g_q \end{bmatrix}, \quad (9-33)$$

which is eq.(7-63) of the book just quoted. Similarly, eqs.(7-64) and (7-65), *loc.cit.*, are special cases of (9-29).

All covariances C_{Pi} and C_{ij} are obtained from the same *covariance function* $C(d)$ which is assumed to depend only on the horizontal distance of the points under consideration:

$$C_{Pi} = C(d_{Pi}), \quad C_{ij} = C(d_{ij}), \quad (9-34)$$

where s_{Pi} is the distance between P and P_i , and s_{ij} is the distance between P_i and P_j .

In the sequel, however, we shall apply the least-squares prediction method not only to homogeneous data such as gravity anomalies, but to the estimation of different quantities of the anomalous gravitational field--such as the disturbing potential, geoidal heights, or spherical-harmonic coefficients--from heterogeneous data--such as gravity anomalies, deflections of the vertical, or satellite data--, arriving at least-squares collocation (sec.11).

10. THE COVARIANCE FUNCTION

As we have just seen, the covariance function of the gravity anomalies is essential for least-squares gravity prediction. Also in the generalization to least-squares collocation, to be considered in the next section, it is necessary to derive all covariances from one basic covariance function $K(P,Q)$, for which we shall take the *covariance function of the disturbing potential* T .

Let $T(P)$ and $T(Q)$ be the disturbing potential T at two points P and Q in space; then $K(P,Q)$ is defined as

$$K(P,Q) = M\{T(P)T(Q)\}, \quad (10-1)$$

where $M\{\cdot\}$ is a suitable averaging operator. This definition is analogous

to formulas such as (9-7), the mean $M(\cdot)$ taking the place of the expectation $E(\cdot)$.

We define the mean M as an average over the whole sphere and over all azimuths, precisely as in (Heiskanen and Moritz, 1967, p.258). Thus we start with the case that both points P and Q lie on the surface of a sphere $r = R$ representing a mean terrestrial sphere which corresponds to the reference ellipsoid as a spherical approximation (p.15). Since the operator M is homogeneous (average over the whole sphere) and isotropic (average over all azimuths), the function $K(P,Q)$ will then be a function only of the spherical distance ψ between P and Q :

$$\begin{aligned} K(P,Q) &= K(\psi) = M\{T(P)T(Q)\} = \\ &= \frac{1}{8\pi^2} \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} \int_{\alpha=0}^{2\pi} T(\theta,\lambda)T(\theta',\lambda') \sin\theta d\theta d\lambda d\alpha. \end{aligned} \quad (10-2)$$

This equation is identical to (7-24), *loc. cit.*, Δg being replaced by T . Here r, θ, λ are spherical coordinates (p.18), the points $P(\theta,\lambda)$ and $Q(\theta',\lambda')$ lie on the sphere $r = R$, and α denotes the azimuth (Fig.10.1). The coordinates (θ',λ') are understood to be related to (θ,λ) by

$$\cos\psi = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\lambda'-\lambda) \quad (10-3)$$

with $\psi = \text{const.}$, but to be arbitrary otherwise.

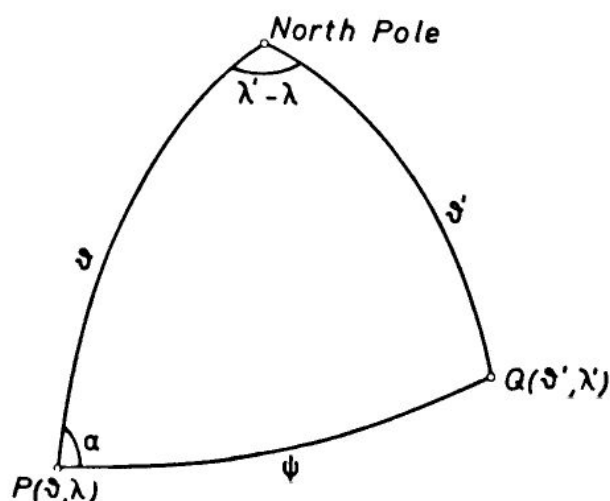


FIGURE 10.1. The basic spherical triangle.

The integration over (θ, λ) expresses homogeneity, and the integration over the azimuth α denotes isotropy. More about this definition of the average M will be found in sec.36.

In agreement with (9-3) we require

$$M\{T\} = 0 . \quad (10-4)$$

Now

$$\begin{aligned} M\{T\} &= \frac{1}{8\pi^2} \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} \int_{\alpha=0}^{2\pi} T(\theta, \lambda) \sin\theta d\theta d\lambda d\alpha \\ &= \frac{1}{8\pi^2} \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} T(\theta, \lambda) \sin\theta d\theta d\lambda \int_0^{2\pi} d\alpha \\ &= \frac{1}{4\pi} \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} T(\theta, \lambda) \sin\theta d\theta d\lambda . \end{aligned} \quad (10-5)$$

This integral is zero if $T(\theta, \lambda)$ does not contain a zero-degree harmonic, which can be achieved by choosing the mass of the reference ellipsoid to be equal to the mass of the earth (Heiskanen and Moritz, 1967, p.99). Similarly, the first degree harmonic term $T_1(\theta, \lambda)$ can be made zero by an appropriate choice of the coordinate origin (*ibid.*, p.100).

Henceforth we shall thus assume that $T(\theta, \lambda)$ contains no spherical harmonics of degrees zero and one; then (10-4) will be satisfied. Thus the spherical-harmonic expansion of T has the form

$$T(\theta, \lambda) = \sum_{n=2}^{\infty} \sum_{m=0}^n \left[\bar{a}_{nm} \bar{R}_{nm}(\theta, \lambda) + \bar{b}_{nm} \bar{S}_{nm}(\theta, \lambda) \right] \quad (10-6)$$

using fully normalized harmonics (p.22).

The spherical harmonic-expansion of the function (10-2) can be written

$$K(\psi) = \sum_{n=2}^{\infty} k_n P_n(\cos\psi) , \quad (10-7)$$

where $P_n(\cos\psi)$ are the (usual or "conventional") Legendre polynomials. The k_n can be expressed in terms of \bar{a}_{nm} and \bar{b}_{nm} by

$$k_n = \sum_{m=0}^n (\bar{a}_{nm}^2 + \bar{b}_{nm}^2) \quad (10-8)$$

(*ibid.*, p.259); note that k_n refers to conventional harmonics, whereas \bar{a}_{nm} and \bar{b}_{nm} are coefficients of fully normalized harmonics.

The extension of the function (10-7) into the space outside the sphere $r = R$ follows uniquely if we require that the function $K(P,Q)$ be harmonic outside this sphere, both as a function of P and as a function of Q . This requirement is evident in view of the definition (10-1) of $K(P,Q)$ as an average product of $T(P)$ and $T(Q)$, since T is assumed to be harmonic outside $r = R$ (p.15).

Now we know (p.20) that the n -th degree spherical harmonic outside a sphere depends on the radius vector r through $r^{-(n+1)}$. Thus $K(P,Q)$ in outer space must have the form

$$K(P,Q) = \sum_{n=0}^{\infty} \frac{\text{constants}}{r^{n+1} r'^{n+1}} P_n(\cos\psi) ,$$

r and r' denoting the radius vectors of P and Q , respectively. On the sphere, for $r = r' = R$, this reduces to (10-7). This determines the constants. The result is

$$K(P,Q) = \sum_{n=2}^{\infty} k_n \left(\frac{R^2}{rr'} \right)^{n+1} P_n(\cos\psi) , \quad (10-9)$$

which expresses the *spatial covariance function of the anomalous potential*.

More about covariance functions will be found in secs. 22, 23, 24, and 34, as well as in (Meissl, 1971a).

11. LEAST-SQUARES COLLOCATION

Consider now least-squares prediction, as discussed in sec.9, for the case that the signal s to be estimated is the anomalous potential $T(P)$ at some point P and that the measurements forming the vector l are arbitrary quantities of the anomalous gravitational field, for instance, gravity anomalies Δg or deflections of the vertical ξ, η . This important problem was first posed in full generality and solved by Krarup (1969).

Any one of these latter quantities may be represented as a *linear functional of the potential* T , for instance, in spherical approximation

$$\Delta g = - \frac{\partial T}{\partial r} - \frac{2}{r} T , \quad (11-1)$$

$$\xi = \frac{1}{\gamma r} \frac{\partial T}{\partial \theta}, \quad \eta = -\frac{1}{\gamma r \sin \theta} \frac{\partial T}{\partial \lambda}, \quad (11-2)$$

r, θ, λ being spherical coordinates (sec.3). Eq.(11-1) is (2-33), and eqs. (11-2) are (2-30) in spherical coordinates. We have written r instead of R since these equations do not necessarily refer to sea level $r = R$ (cf. also sec.42); and γ is normal gravity, in keeping with (2-31).

Generally we have

$$l_i = L_i T \quad (11-3)$$

or

$$l = BT \quad (11-4)$$

where the "vector" B comprises the functionals L_i :

$$B = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_q \end{bmatrix}, \quad (11-5)$$

Thus the problem is to find T if q linear functionals $L_i T$ are given by measurement. The determination of a function by fitting an analytical approximation to a certain number of given linear functionals is called collocation and is frequently used in numerical mathematics; cf. (Collatz, 1966, p.29). Therefore, the present method for determining the gravitational field by least-squares prediction is called *least-squares collocation*.

The application of (9-28) to the present problem gives at once

$$\hat{T}(P) = \begin{bmatrix} C_{P1} & C_{P2} & \dots & C_{Pq} \end{bmatrix} \begin{bmatrix} C_{11} & C_{12} & \dots & C_{1q} \\ C_{21} & C_{22} & \dots & C_{2q} \\ \vdots & \vdots & & \vdots \\ C_{q1} & C_{q2} & \dots & C_{qq} \end{bmatrix}^{-1} \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_q \end{bmatrix}. \quad (11-6)$$

This is analogous to the interpolation formula (9-33), but the covariances are different.

Before computing these covariances, let us verify that our quantities T and l_i are centered. The equivalent of the second equation of (9-3) for the present case is

$$M\{T\} = 0, \quad (11-7)$$

which is (10-4) and means that T does not contain a zero-degree spherical harmonic. The first equation of (9-3) becomes

$$M\{l_i\} = 0. \quad (11-8)$$

Using (11-3) we may write

$$M\{l_i\} = M\{L_i T\} = L_i M\{T\}, \quad (11-9)$$

which is zero by (11-7), so that (11-8) is satisfied. Here we have assumed --as we shall always assume in the sequel--that the averaging operator M and the linear functionals L_i commute, so that the order of these two operations can be interchanged. This will be justified later (sec.36).

Covariance propagation. In (11-6) we have

$$C_{Pi} = \text{cov}(T(P), l_i) = M\{T(P)l_i\}, \quad (11-10)$$

$$C_{ij} = \text{cov}(l_i, l_j) = M\{l_i l_j\}. \quad (11-11)$$

We may write (11-3) in the form

$$l_i = L_i^Q T(Q), \quad (11-12)$$

indicating that the functional L_i is applied to T as a function of the independent point variable Q . Then (11-10) becomes

$$C_{Pi} = M\{T(P)L_i^Q T(Q)\} = L_i^Q M\{T(P)T(Q)\},$$

on using the commutativity of M and L_i . By the definition (10-1) we thus obtain

$$C_{Pi} = L_i^Q K(P, Q). \quad (11-13)$$

similarly we get

$$C_{ij} = M \{ L_i^P T(P) L_j^Q T(Q) \} = L_i^P L_j^Q M \{ T(P) T(Q) \} ,$$

so that

$$C_{ij} = L_i^P L_j^Q K(P, Q) . \quad (11-14)$$

Equations (11-13) and (11-14) express the required covariances in terms of the basic covariance function $K(P, Q)$ of T , as defined in the preceding section. They show how the covariances "propagate" through linear operations L_i ; they may thus be called formulas of *covariance propagation*.

In fact, these formulas are, so to speak, the continuous analogue of the usual matrix formulas for error propagation, or covariance propagation, in adjustment computations. Let s be a vector and let K be its covariance matrix

$$K = \text{cov}(s, s) . \quad (11-15)$$

Consider another vector l which is a linear function of s :

$$l = Bs , \quad (11-16)$$

B being an appropriate matrix. Then, by the usual covariance propagation,

$$\text{cov}(s, l) = KB^T , \quad (11-17)$$

$$\text{cov}(l, l) = BKB^T , \quad (11-18)$$

which corresponds to (11-13) and (11-14). Later we shall exploit these analogies even more fully (secs. 21 and 25).

Let us now return to (11-13) and (11-14). The exact meaning of these expressions is as follows. Eq. (11-13) states that the linear functional L_i is applied to $K(P, Q)$, considered as a function of Q only, P being regarded as a constant parameter. In (11-14) we first apply L_j to $K(P, Q)$, considered as a function of Q . The result, $L_j^Q K(P, Q)$, then depends only on P . It may be regarded as a function of P , to which the operator L_i is finally applied; this gives C_{ij} .

Example. Let us illustrate this procedure by means of an example. Let

$$l_1 = L_1 T = \gamma \xi(P_1) = \left(\frac{1}{r} \frac{\partial T}{\partial \theta} \right)_{P_1}, \quad (11-19)$$

$$l_2 = L_2 T = \gamma \eta(P_2) = - \left(\frac{1}{r \sin \theta} \frac{\partial T}{\partial \lambda} \right)_{P_2}, \quad (11-20)$$

the linear functionals L_1 and L_2 being given by (11-2), and P_1 and P_2 being fixed points. By (11-14),

$$\text{cov}[\gamma \xi(P_1), \gamma \eta(P_2)] = \text{cov}(l_1, l_2) = C_{12} = L_1^P L_2^Q K(P, Q).$$

Now, with $P = (r, \theta, \lambda)$ and $Q = (r', \theta', \lambda')$,

$$L_2^Q K(P, Q) = - \frac{1}{r' \sin \theta'} \frac{\partial K(P, Q)}{\partial \lambda'}$$

for $Q = P_2$, and

$$C_{12} = L_1 \left[- \frac{1}{r' \sin \theta'} \frac{\partial K(P, Q)}{\partial \lambda'} \right] = - \frac{1}{r r' \sin \theta'} \frac{\partial^2 K(P, Q)}{\partial \theta \partial \lambda'} \quad (11-21)$$

for $P = P_1$, $Q = P_2$. By (10-9), the function $K(P, Q)$ depends explicitly on r and r' and implicitly on θ, λ and θ', λ' through (10-3). Therefore, the differentiations in (11-21) can be performed without difficulty.

Estimation of functionals. Instead of the potential $T(P)$, let us directly estimate another signal, that is, another element s of the anomalous gravitational field, for instance a geoidal height or a spherical-harmonic coefficient of T . We write

$$s = ST = S^P T(P), \quad (11-22)$$

since any element of the anomalous gravitational field can be expressed as a linear functional S of T .

The application of S^P to (11-6) gives

$$\xi = \begin{bmatrix} c_{s1} & c_{s2} & \dots & c_{sq} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1q} \\ c_{21} & c_{22} & \dots & c_{2q} \\ \vdots & \vdots & & \vdots \\ c_{q1} & c_{q2} & \dots & c_{qq} \end{bmatrix}^{-1} \begin{bmatrix} 1_1 \\ 1_2 \\ \vdots \\ 1_q \end{bmatrix}, \quad (11-23)$$

where

$$c_{si} = S^P c_{pi}, \quad (11-24)$$

the remainder c_{11}^{-1} being unchanged. In fact, only the first row vector depends on the variable P , on which the functional $S = S^P$ acts.

Let us now estimate m signals forming the vector

$$s = \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{bmatrix}, \quad (11-25)$$

where

$$s_k = S_k T = S_k^P T(P). \quad (11-26)$$

For each signal s_k we get an equation (11-23), which can be combined into one matrix equation

$$\xi = c_{s1} c_{11}^{-1}, \quad (11-27)$$

where

$$c_{s1} = \begin{bmatrix} S_1^P c_{p1} & S_1^P c_{p2} & \dots & S_1^P c_{pq} \\ S_2^P c_{p1} & S_2^P c_{p2} & \dots & S_2^P c_{pq} \\ \vdots & \vdots & & \vdots \\ S_m^P c_{p1} & S_m^P c_{p2} & \dots & S_m^P c_{pq} \end{bmatrix}. \quad (11-28)$$

Any element of this matrix is given by

$$S_k^P C_{P1} = S_k^P L_1^Q K(P, Q) , \quad (11-29)$$

in view of (11-13), but this is precisely

$$\text{cov}(s_k, l_1) \quad (11-30)$$

as computed by the covariance propagation law (11-14) from (11-12) and (11-26)! This means that we can directly apply the prediction formula (9-28) to the estimation of the vector s , after computing the covariance matrices C_{s1} and C_{11} by covariance propagation from $K(P, Q)$. The result is the same as by first computing \hat{T} from (11-6) and then

$$\hat{s}_k = S_k \hat{T} . \quad (11-31)$$

The collocation formulas (11-6) for the function T and (11-27) for directly computing any functionals \hat{s} of T are completely consistent.

In other terms, we may apply the basic least-squares prediction formula (9-28) for the computation of any element of the anomalous gravitational field from data which are arbitrary other elements of this field, if we compute all covariances by covariance propagation from the same function $K(P, Q)$; the results so obtained will be consistent. The covariances are thus seen to carry the analytical structure of the anomalous gravitational field: they must be rigorously derived from one covariance function $K(P, Q)$ by formulas such as (11-13) and (11-14).

Conversely, (11-6) may be considered as a special case of the estimation formula (11-23) for a linear functional, if we take S to be the "evaluation functional"

$$ST = T(P) , \quad (11-32)$$

associating to the function T its value at the point p .

Accuracy. The accuracy of the least-squares collocation formula (11-27) is given by the error covariance matrix E_{ss} (9-29), in which C_{11} and C_{s1} are the same matrices as in (11-27) and

$$C_{1s} = C_{s1}^T . \quad (11-33)$$

The $m \times m$ matrix C_{ss} is computed by covariance propagation for the functional (11-22); eq.(11-14) shows immediately that C_{ss} has in the h -th row and k -th column the element

$$(C_{ss})_{hk} = S_h^P S_k^Q K(P,Q) , \quad h,k = 1, 2, \dots, m . \quad (11-34)$$

The diagonal elements $(E_{ss})_{kk}$ of the error covariance matrix give the error variance

$$\sigma_k^2 = M \left\{ \epsilon_k^2 \right\} = M \left\{ (\hat{s}_k - s_k)^2 \right\} \quad (11-35)$$

of the estimated signal \hat{s} ; cf. (9-16) with E replaced by M .

The precise meaning of this average M should be clearly kept in mind. As we have seen in sec.10, M is a homogeneous and isotropic average over the sphere, and σ_k^2 is a mean square estimation error in the sense of this average. This definition appears to be the natural one for characterizing accuracies on a global scale.

12. INVARIANCE PROPERTIES; ANALYTICAL COLLOCATION

The least-squares collocation solution (11-27) possesses a series of important structural properties, related to the fact that all estimated quantities refer to the same gravitational field.

Reproduction of data. If the estimate potential \hat{T} is to be consistent with the data, then the function \hat{T} must reproduce the given functionals l_i , that is, there must be

$$l_i = L_i T = L_i \hat{T} = \hat{l}_i . \quad (12-1)$$

This is easily proved directly. If the vector s to be estimated by (11-27) coincides with the q -vector l itself, then $C_{s1} = C_{11}$, and (11-27) gives

$$\hat{l} = C_{11} C_{11}^{-1} l = l , \quad (12-2)$$

which proves (12-1).

Invariance with respect to linear transformations. In the preceding section we have already seen that least-squares collocation is *invariant with respect to any linear transformation of the estimated signal*: one obtains the same result by first estimating the potential \hat{T} by (11-6) and then computing a linear functional $\xi_k = S_k \hat{T}$, or by directly estimating ξ_k by (11-27), provided the covariances are properly derived.

The method is also *invariant with respect to linear transformations of the data*. Instead of the q data l_i forming the vector l , let us introduce other q data l'_i forming the vector l' , which are related to l_i by the transformation

$$l' = Al, \quad l = A^{-1}l', \quad (12-3)$$

A being a regular (invertible) $q \times q$ matrix. Then

$$C_{s1'} = M\{s1'^T\} = M\{s1^T A^T\} = M\{s1^T\} A^T,$$

or

$$C_{s1'} = C_{s1} A^T, \quad (12-4)$$

and similarly

$$C_{1'1'} = AC_{11}A^T. \quad (12-5)$$

Applying (11-27) to l' we have

$$\xi = C_{s1'} C_{1'1'}^{-1} l'. \quad (12-6)$$

By (12-3), (12-4), and (12-5) this becomes

$$\xi = C_{s1} A^T (AC_{11}A^T)^{-1} Al = C_{s1} A^T (A^T)^{-1} C_{11}^{-1} A^{-1} Al,$$

so that

$$\xi = C_{s1} C_{1'1'}^{-1} l' = C_{s1} C_{11}^{-1} l, \quad (12-7)$$

which was to be shown.

Analogous invariance properties are well known from least-squares adjustment; cf. (Tienstra, 1956, p.154).

Analytical collocation. What happens if the function $K(P,Q)$, from which the covariances C_{P1} and C_{1j} in (11-6) are derived, is not the covariance function of T , but an arbitrary analytical function of form (10-9), with nonnegative coefficients k_n satisfying only the condition that the infinite series¹ converges for $r, r' \geq R$? Such a function $K(P,Q)$ is symmetric in P and Q and harmonic as a function of both P and Q , outside and on the sphere of radius R ; it is called a *kernel function*.

In this case, the complete analytical structure of the anomalous gravitational field is preserved: we get a completely consistent gravitational field in the sense that the given data l_i are exactly reproduced, and the various estimated quantities s_k refer to the same field, in view of the invariance with respect to linear transformations of the signal.

In other terms, the given q data l_i do not completely determine the potential T . There are infinitely many functions $\hat{T}(P)$ compatible with the given data, and the use of different kernel functions $K(P,Q)$ in (11-6) corresponds to different possible gravitational fields, all of which are internally consistent.

On putting

$$b = C_{11}^{-1} l, \quad (12-8)$$

so that b is a q -vector, we see that (11-6) has the form

$$\hat{T}(P) = \sum_{i=1}^q b_i C_{P1} \quad (12-9)$$

By (11-13),

$$C_{P1} = L_1^Q K(P,Q), \quad (12-10)$$

which are functions $\phi_i(P)$ of P . Therefore we may write (12-9) as

$$\hat{T}(P) = \sum_{i=1}^q b_i \phi_i(P), \quad (12-11)$$

¹ It must be a true infinite series containing infinitely many nonvanishing coefficients k_n ; otherwise the covariance matrix C_{11} derived from it may not be invertible.

as a linear combination of q "base functions" $\phi_i(P)$.

If the kernel function $K(P,Q)$ has a simple analytical expression, then the base functions

$$\phi_i(P) = L_i^Q K(P,Q) \quad (12-12)$$

will also be simple analytical functions, so that (12-9) may be considered as an analytical approximation to the function $T(P)$ by means of a linear combination of q base functions. This aspect of collocation as an analytical approximation method is described in detail in (Moritz, 1978b).

Analytical collocation methods, using a general kernel function, have been considered, e.g., by Krarup (1978), Lelgemann (1978), and Tscherning (1978b).

Of course, only if $K(P,Q)$ is the covariance function will the estimates (11-6) or (11-27) have the property of minimum variance. The error variances and covariances, forming the error covariance matrix $C_{\epsilon\epsilon} = E_{ss}$, for the case of a general kernel function can be obtained from (9-21). Let us denote the "true" covariance function by $K(P,Q)$, and the analytical kernel function used in the collocation procedure by $\bar{K}(P,Q)$. The matrices \bar{C}_{s1} and \bar{C}_{11} are derived from $\bar{K}(P,Q)$ in the same way as C_{s1} and C_{11} are derived from $K(P,Q)$. Then the analytical collocation gives

$$\hat{s} = \bar{C}_{s1} \bar{C}_{11}^{-1} l, \quad (12-13)$$

so that, in (9-21), we must put

$$H = \bar{C}_{s1} \bar{C}_{11}^{-1}, \quad (12-14)$$

obtaining

$$E_{ss} = C_{ss} - C_{s1} C_{11}^{-1} C_{1s} + (\bar{C}_{s1} \bar{C}_{11}^{-1} - C_{s1} C_{11}^{-1}) C_{11} (\bar{C}_{s1} \bar{C}_{11}^{-1} - C_{s1} C_{11}^{-1})^T. \quad (12-15)$$

For $\bar{K}(P,Q) = K(P,Q)$ this coincides with the optimum estimate (9-29).

This analytical aspect of least-squares collocation is of basic theoretical and practical significance. It shows that a general kernel function can be used to obtain a consistent gravitational field which is compatible with the given measurements. The covariance function $K(P,Q)$ cannot be exactly determined empirically since for this purpose we need the function

$T(\theta, \lambda)$ entering in (10-2). That is, we should know the anomalous potential everywhere on the sphere $r = R$, which obviously is not the case. The empirical covariance function used will be an analytical expression fitted to the given data (sec.23) and will thus have the character of a general kernel function.

13. APPLICATION TO BJERHAMMAR'S PROBLEM

As an illustration, let us apply collocation to the following problem. Let the gravity anomalies $\Delta g_1, \Delta g_2, \dots, \Delta g_n$ at n points on the earth's surface, at elevations h_1, h_2, \dots, h_n , be known from observation; to determine a gravity field consistent with these observations.

This formulation of the main problem of gravimetric geodesy, due to (Bjerhammar, 1964), is in a way more realistic than the usual formulation in terms of a boundary-value problem, since we observe at discrete points only. In fact, Bjerhammar's problem has contributed, in several ways, to a clarification of the conceptual foundations of physical geodesy.

In the present solution we shall use an analytical continuation of the gravity anomaly Δg down to sea level. For the usual case, in which the measured gravity anomalies are assumed to be known at every point of the earth's surface, this method has been described in (Heiskanen and Moritz, 1967, section 8-10); then, however, the problem may fail to have a rigorous solution because of possible singularities in downward continuation (*ibid.*, p.321). This difficulty does not arise in the present case of discrete observations, as we shall see.

We shall use our customary spherical approximation, which amounts to tolerating a relative error of the order of the flattening $f = 0,3\%$ in quantities of the anomalous gravity field. This permits us to formally replace the reference ellipsoid by a sphere. Then the free-air anomalies downward continued to sea level, denoted by Δg^* , can be considered as a function $\Delta g^*(\theta, \lambda)$ on the surface of a terrestrial sphere (of radius R), θ and λ being the spherical coordinates introduced in section 4.

Thus the problem is to compute this function $\Delta g^*(\theta, \lambda)$ on the sphere from gravity anomalies $\Delta g_1, \Delta g_2, \dots, \Delta g_q$ measured at the earth's surface, that is, above the sphere. If we put

$$s = \Delta g_p^* \tag{13-1}$$

at some point $P(R, \theta, \lambda)$ and

$$l_i = \Delta g_i \quad (i = 1, 2, \dots, q) \quad (13-2)$$

we may apply the basic formula (11-23); there remains the computation of the covariances.

Since the external potential together with its analytical continuation forms a single harmonic function (sec.6), Δg_p^* and Δg_i must be values of the same spatial analytic function $\Delta g(r, \theta, \lambda)$:

$$\Delta g_p^* = \Delta g(R, \theta, \lambda), \quad (13-3)$$

$$\Delta g_i = \Delta g(r_i, \theta_i, \lambda_i) \quad \text{with} \quad r_i = R + h_i. \quad (13-4)$$

Therefore, all occurring covariances, which are covariances between Δg_p^* and Δg_i or between Δg_i and Δg_j , will be values of the same function $C(P, Q)$, the spatial covariance function of the gravity anomaly.

For a point $P(r, \theta, \lambda)$ outside or on the sphere $r = R$ we have the spherical-harmonic expansion

$$T(P) = \sum_{n=2}^{\infty} \left(\frac{R}{r} \right)^{n+1} T_n(\theta, \lambda), \quad (13-5)$$

$T_n(\theta, \lambda)$ being a Laplace surface harmonic of degree n . The corresponding expansion for the spatial gravity anomaly is

$$\Delta g(P) = \frac{1}{r} \sum_{n=2}^{\infty} (n-1) \left(\frac{R}{r} \right)^{n+1} T_n(\theta, \lambda). \quad (13-6)$$

(Heiskanen and Moritz, 1967, pp.88-89). Therefore, the linear operation transforming T into Δg consists of multiplication of the n -th degree harmonic by $(n-1)/r$.

The covariance function of T is

$$K(P, Q) = \sum_{n=2}^{\infty} k_n \left(\frac{R^2}{rr'} \right)^{n+1} P_n(\cos \psi). \quad (13-7)$$

The corresponding expansion for the covariance function $C(P, Q)$ of Δg is obtained by multiplication of the n -th degree harmonic by

$$\frac{n-1}{r} \frac{n-1}{r'} = \frac{(n-1)^2}{rr'} ,$$

in agreement with the covariance propagation formula (11-14). Thus

$$C(P,Q) = \sum_{n=2}^{\infty} (n-1)^2 k_n \frac{R^{2n+2}}{(rr')^{n+2}} P_n(\cos\psi) , \quad (13-8)$$

or

$$C(P,Q) = \sum_{n=2}^{\infty} c_n \left(\frac{R^2}{rr'} \right)^{n+2} P_n(\cos\psi) \quad (13-9)$$

where

$$c_n = \left(\frac{n-1}{R} \right)^2 k_n . \quad (13-10)$$

Therefore we must put in (11-23):

$$C_{si} = \text{cov}(\Delta g_p^*, \Delta g_i) = \sum_{n=2}^{\infty} c_n \left(\frac{R}{r_i} \right)^{n+2} P_n(\cos\psi_{pi}) , \quad (13-11)$$

$$C_{ij} = \text{cov}(\Delta g_i, \Delta g_j) = \sum_{n=2}^{\infty} c_n \left(\frac{R^2}{r_i r_j} \right)^{n+2} P_n(\cos\psi_{ij}) , \quad (13-12)$$

since $r_p = R$.

Since the function $K(P,Q)$ is assumed to be regular outside and on the sphere $r = R$, the same holds for $C(P,Q)$ and, therefore, for the computed downward continuation $\Delta g^*(\theta, \lambda)$. Thus collocation ensures, in fact, a smooth regular downward continuation of Δg (provided we have only a finite number of given Δg_i). As the point P can be an arbitrary point on the sphere $r = R$ (Fig.13.1), the collocation solution automatically combines *downward continuation and interpolation* in a natural way, so as to obtain a smooth solution.

As we shall see, smoothness can even be defined in a precise way, in terms of a norm which is minimized (sec.25).

The fact that a smooth analytical downward continuation is, so to speak, built in with this method, is of great importance since downward continuation is, in general, a difficult and unstable operation. The same principle can, for instance, be applied for the downward continuation of aerial gravity measurements (Moritz, 1970, sec.6).

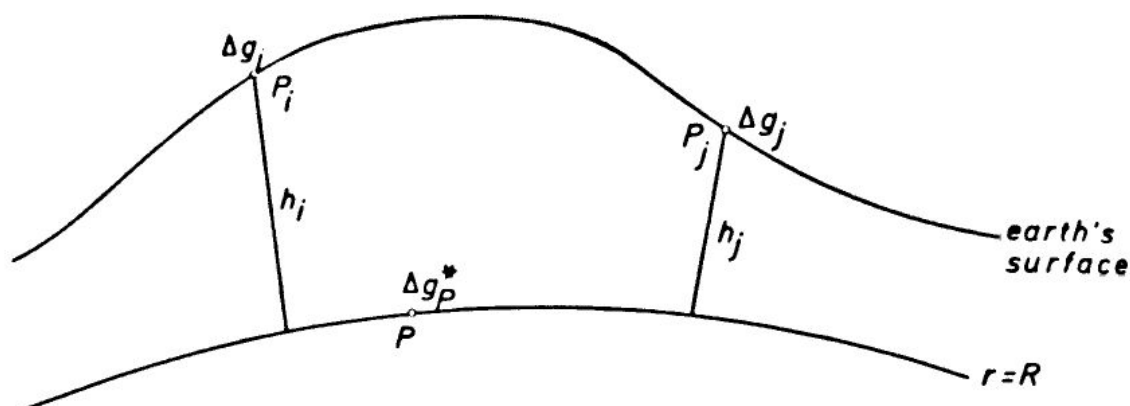


FIGURE 13.1. Downward continuation of Δg .

Invariance. By computing Δg_P^* at every point of the sphere $r = R$, we get a function $\Delta g^*(\theta, \lambda)$. To this function we may then apply the Stokes-Pizzetti formula

$$T = \frac{R}{4\pi} \iint_{\sigma} \Delta g^* S(r, \psi) d\sigma \quad (13-13)$$

to obtain T at any point on the earth's surface or in outer space (Heiskanen and Moritz, 1967, p.319).

We may, however, also compute T at a point $P(r, \theta, \lambda)$ directly from (11-6), with C_{ij} as before and C_{Pi} given by

$$C_{Pi} = \text{cov}[T(P), \Delta g_i] = \sum_{n=2}^{\infty} \frac{n-1}{r_i} k_n \left(\frac{R^2}{rr_i} \right)^{n+1} P_n(\cos \psi), \quad (13-14)$$

which follows from (13-7) by multiplying each term by $(n-1)/r'$ and then substituting $r' = r_i$; this is a direct consequence of (11-13). This is even simpler than going through Δg^* and (13-13), but the result is identical, in view of the invariance of least-squares collocation with respect to linear transformations; T being a linear functional of Δg^* .

Significance of Runge's theorem. The present approach presupposes the possibility of analytical continuation of the external potential T down to sea level as represented by the sphere $r = R$. This is possible to as good an approximation as we like, in view of Runge's theorem (sec.8). This justifies the assumption that T is a regular harmonic function on and outside the sphere $r = R$; this assumption will be used throughout.

14. COLLOCATION WITH RANDOM ERRORS

If the measurements l_i forming the q -vector l are affected by random measuring errors n_i , then instead of (11-3) we have

$$l_i = L_i T + n_i, \quad i = 1, 2, \dots, q. \quad (14-1)$$

Writing

$$L_i T = t_i \quad (14-2)$$

we get

$$l_i = t_i + n_i. \quad (14-3)$$

In this way we have decomposed the observation l_i into a "signal" t_i and the "noise" n_i . The "signal part" of l_i represents the gravitational field element $L_i T$ of which l_i is the measurement, and the "noise" is a synonym for the random measuring errors. The terminology, signal and noise, comes from communication engineering in which statistical prediction techniques are widely used.

In symbolic notation these equations are written:

$$l = BT + n, \quad (14-4)$$

$$BT = t, \quad (14-5)$$

$$l = t + n, \quad (14-6)$$

generalizing (11-4).

The noise n is a genuine random (stochastic) quantity. It possesses a probability distribution with a mathematical expectation denoted by E . In contrast to this, the statistical treatment of s has a more formal character. The operator M denotes a homogeneous and isotropic average over the sphere, defined by (10-2), rather than an expectation in a probabilistic sense: M describes the average global behavior of the anomalous gravitational field.

As we have seen,

$$M(t) = 0$$

(14-7)

for all signals t , that is, all elements of the gravitational field, provided the reference field is chosen so that the anomalous potential τ does not contain a zero-degree harmonic: from (11-7) we get

$$M(t) = M(BT) = BM(T) = 0.$$

For the same reason we have, of course, for the signals (11-22) to be estimated:

$$M(s) = 0.$$

(14-8)

On the other hand, the signals are not stochastic quantities in the same sense as the noise n : repeated observations of the same quantity give different n but s remains the same. Thus, the expectation E does not affect the signals, whence

$$E(s) = s, \quad E(t) = t.$$

(14-9)

However, for the noise we have

$$E(n) = 0$$

(14-10)

since random measuring errors have zero expectation by definition.

The operation of the spherical average M on the noise has not yet been defined. The simplest and most suitable definition is that M leaves n unchanged:

$$M(n) = n.$$

(14-11)

This definition is analogous to (14-9); a detailed justification will be found in sec.36.

Now we define the *total average* \bar{E} as

$$\bar{E} = EM;$$

(14-12)

that is, we average both over the probabilistic distribution of n and over the sphere. With this definition we have

$$\bar{E}(s) = EM(s) = 0$$

(14-13)

by (14-8), and likewise

$$\bar{E}(t) = 0 . \quad (14-14)$$

But also

$$\bar{E}(n) = EM(n) = E(n) = 0 , \quad (14-15)$$

in view of (14-11) and (14-10). Hence

$$\bar{E}(l) = \bar{E}(t) + \bar{E}(n) = 0 , \quad (14-16)$$

by (14-6), so that all our quantities are centered.

Let us now consider the covariances corresponding to the new average \bar{E} . For the signal covariance matrices we have

$$C_{ss} = \bar{E}\{ss^T\} = EM\{ss^T\} = E\{M\{ss^T\}\} = M\{ss^T\} \quad (14-17)$$

since E does not affect signals. Similarly,

$$C_{tt} = \bar{E}\{tt^T\} = M\{tt^T\} . \quad (14-18)$$

For the covariance matrix C_{nn} of the noise we find

$$C_{nn} = \bar{E}\{nn^T\} = EM\{nn^T\} = E\{nn^T\} . \quad (14-19)$$

The mixed covariances are zero:

$$C_{tn} = \bar{E}\{tn^T\} = EM\{tn^T\} = M\{t\}E\{n^T\} = 0 , \quad (14-20)$$

that is, under the average \bar{E} , the signal and the noise are uncorrelated.

In order to apply the basic prediction formula (9-28) we need C_{s1} and C_{11} . Now,

$$\begin{aligned} C_{11} &= \bar{E}\{ll^T\} = \bar{E}\{(t+n)(t^T+n^T)\} \\ &= \bar{E}\{tt^T\} + \bar{E}\{tn^T\} + \bar{E}\{nt^T\} + \bar{E}\{nn^T\} \\ &= C_{tt} + C_{tn} + C_{nt} + C_{nn} . \end{aligned} \quad (14-21)$$

By (14-20) we also have

$$C_{nt} = C_{tn}^T = 0 ,$$

(14-22)

and hence

$$C_{11} = C_{tt} + C_{nn} .$$

(14-23)

Thus C_{11} is simply the sum of the covariance matrices of signal and noise. This is rather remarkable since the signal covariance matrix is an average $M\{tt^T\}$, and the noise covariance matrix is an expectation $E\{nn^T\}$.

For the cross-covariance matrix C_{s1} we find

$$C_{s1} = E\{s1^T\} = E\{s(t^T + n^T)\}$$

$$= E\{st^T\} + E\{sn^T\} = C_{st} + C_{sn} .$$

(14-24)

Since signal and noise are uncorrelated, we have

$$C_{sn} = 0 ,$$

(14-25)

and

$$C_{st} = E\{st^T\} = EM\{st^T\} = M\{st^T\} ,$$

both s and t being signals (elements of the anomalous gravitational field).

Thus

$$C_{s1} = C_{st}$$

(14-26)

is a pure signal covariance matrix.

In view of (14-23) and (14-26), the prediction formula (9-28) becomes

$$\xi = C_{st}(C_{tt} + C_{nn})^{-1} .$$

(14-27)

This is the fundamental formula for *least-squares collocation with noise*.

The signal covariances are to be derived from the basic covariance function $K(P,Q)$. Corresponding to

$$t_i = L_i^T, \quad s_k = S_k^T \quad (14-28)$$

(these are equations (14-2) and (11-26)) we find for the elements of C_{tt} and C_{st} :

$$(C_{tt})_{ij} = L_i^P L_j^Q K(P,Q), \quad (14-29)$$

$$(C_{st})_{ki} = S_k^P L_i^Q K(P,Q). \quad (14-30)$$

The comparison with (11-14) and (11-29) shows that, in the absence of measuring errors, we have

$$C_{tt} = C_{11}, \quad C_{st} = C_{s1} \quad \text{if} \quad n = 0, \quad (14-31)$$

as it must be.

Let us thus compare the collocation formulas with-random errors, (14-27), and without errors, (11-27), which may be written in the present notation as

$$\hat{s} = C_{st} C_{tt}^{-1} t. \quad (14-32)$$

We see that the only difference is the presence of C_{nn} in (14-27). This is the covariance matrix of the observational errors, defined by (14-19). (As a matter of fact, noise covariance matrices C_{nn} are the usual variance-covariance matrices in adjustment computations.)

Let us now apply (14-27) to the estimation of the potential T . Then

$$s = T(P), \quad (14-33)$$

and

$$C_{st} = \begin{bmatrix} C_{P1} & C_{P2} & \dots & C_{Pq} \end{bmatrix} \quad (14-34)$$

where the C_{Pi} are given by (11-13). Thus (14-27) becomes

$$\hat{T}(P) = [C_{P1} \ C_{P2} \ \dots \ C_{Pq}] \begin{bmatrix} C_{11}+D_{11} & C_{12}+D_{12} & \dots & C_{1q}+D_{1q} \\ C_{21}+D_{21} & C_{22}+D_{22} & \dots & C_{2q}+D_{2q} \\ \vdots & \vdots & & \vdots \\ C_{q1}+D_{q1} & C_{q2}+D_{q2} & \dots & C_{qq}+D_{qq} \end{bmatrix}^{-1} \begin{bmatrix} 1_1 \\ 1_2 \\ \vdots \\ 1_q \end{bmatrix}, \quad (14-35)$$

which is the generalization of (11-6) to the case of random errors. The only difference is the presence of noise covariances; we have put

$$C_{nn} = D = [D_{ij}]. \quad (14-36)$$

The other covariances, C_{Pi} and C_{ij} , are the same in (14-35) and (11-6).

The fact, mentioned above, that the matrix C_{st} is a pure signal covariance matrix, is of great importance. To see this, put

$$b = (C_{tt} + C_{nn})^{-1}1 \quad (14-37)$$

and write (14-35) in the form

$$\hat{T}(P) = \sum_{i=1}^q b_i C_{Pi}. \quad (14-38)$$

Exactly as in the errorless case (12-9), \hat{T} is expressed as a linear combination of pure signal covariance functions, which are to be considered as analytical base functions (12-12):

$$\hat{T}(P) = \sum_{i=1}^q b_i \phi_i(P). \quad (14-39)$$

The transition from T to other quantities s_k of the anomalous gravitational field--such as geoidal heights, deflections of the vertical, or gravity anomalies--is effected by linear functional operations (11-26). Thus (14-39) gives

$$s_k = S_k \hat{T} = \sum_{i=1}^q b_i S_k \phi_i(P). \quad (14-40)$$

The functional S_k acts analytically on the base functions ϕ_i . Since,

by (11-13),

$$S_k \phi_i(P) = S_k C_{Pi} = S_k^P L_i^Q K(P, Q), \quad (14-41)$$

this is nothing else than covariance propagation, which is again seen to carry the precise mathematical structure of the gravitational field.

The coefficients b_k , given by the vector (14-37), remain the same for all signals (T or s_k); they depend on the noise covariance matrix C_{nn} and are determined in such a way that the effect of measuring errors n is minimized.

Thus the noise affects only the determination of the coefficients b_k but not the base functions ϕ_k : there is no danger that statistics spoils the analytical field structure.

The error covariance matrix E_{ss} for the estimate (14-27) is obtained from (9-29):

$$E_{ss} = C_{ss} - C_{st}(C_{tt} + C_{nn})^{-1}C_{ts}; \quad (14-42)$$

the matrix C_{ss} is given by (11-34).

Filtering and prediction. Let us apply (14-27) to the signals t , which are the signal parts of the observations l as expressed by (14-6). With $s = t$ we get

$$\hat{t} = C_{tt}(C_{tt} + C_{nn})^{-1}l. \quad (14-43)$$

This equation gives the optimal estimate for the signal part of the observations l themselves; the noise n has been filtered out in the best possible way. Thus (14-43) may be said to describe the *filtering* of the observations l . There is no prediction in this case since, besides the observation signals t , no new signals are computed.

Let us now solve (14-43) for l :

$$l = (C_{tt} + C_{nn})C_{tt}^{-1}\hat{t} \quad (14-44)$$

and substitute into (14-27), whence

$$\hat{s} = C_{st}C_{tt}^{-1}\hat{t}. \quad (14-45)$$

This equation predicts new signals s on the basis of the filtered observations (14-43).

It is quite remarkable that this equation has the form (14-32) of *errorless* collocation, with l replaced by \hat{t} . Thus we have split up least squares collocation with random errors into two consecutive steps:

1. Filtering of the data by (14-43)
2. Application of the errorless collocation formula (14-45) to the filtered data.

Briefly we may say that we have split up collocation with noise into filtering and pure prediction. This interpretation shows again that collocation with noise has the same analytical structure (expressed by step 2) as errorless collocation.

The decomposition of (14-27) into (14-43) and (14-45) has theoretical rather than practical importance since its numerical execution would involve the inversion of two different covariance matrices $C_{tt} + C_{nn}$ and C_{tt} , whereas in the direct solution (14-27) the inversion of the first matrix suffices for all signals s .

15. APPLICATION TO GEOID DETERMINATION

To illustrate the general formulas, let us consider an example which, though being somewhat simplified, shows the essential features of the method and also has practical significance.

We take the determination of the geoidal height N from gravity anomalies Δg and deflections of the vertical components ξ and η . We assume the usual simplified situation underlying Stokes' formula: all quantities refer to sea level, there are no masses outside the geoid, and the reference ellipsoid is formally treated as a sphere (spherical approximation; cf. sec.2). The components ξ and η have been determined by the astrogeodetic method and refer to a geocentric ellipsoid, so that there are no first-degree spherical harmonics. Furthermore, we take f observation points and assume that Δg , ξ , η are given at each of these points, so that there are $q = 3f$ observations.

In the basic formula (14-27),

$$\mathfrak{s} = C_{st}(C_{tt} + C_{nn})^{-1}l, \quad (15-1)$$

we thus have

$$s = N(P) , \quad (15-2)$$

the geoidal height at some point P , and

$$l = \begin{bmatrix} \Delta g_1 \\ \vdots \\ \Delta g_f \\ \xi_1 \\ \vdots \\ \xi_f \\ \eta_1 \\ \vdots \\ \eta_f \end{bmatrix} . \quad (15-3)$$

Let us assume that all Δg have been observed with the same standard error m_g , all ξ with the same standard error m_ξ , and all η with the same standard error m_η ; all measurements are supposed to be uncorrelated. Then the covariance matrix C_{nn} of the noise is the diagonal matrix

$$C_{nn} = \begin{bmatrix} m_g^2 & & & & & \\ & \ddots & & & & \\ & & m_g^2 & & & \\ & & & m_\xi^2 & & \\ & & & & \ddots & \\ & & & & & m_\xi^2 \\ & & & & & & m_\eta^2 \\ & & & & & & & \ddots \\ & & & & & & & & m_\eta^2 \end{bmatrix} . \quad (15-4)$$

The signal covariance matrices are

$$C_{st} = \begin{bmatrix} C_{P1}^{Ng} & \dots & C_{Pf}^{Ng} & C_{P1}^{N\xi} & \dots & C_{Pf}^{N\xi} & C_{P1}^{N\eta} & \dots & C_{Pf}^{N\eta} \end{bmatrix} , \quad (15-5)$$

$$C_{tt} = \begin{bmatrix} C_{11}^{gg} & \dots & C_{1f}^{gg} & C_{11}^{g\xi} & \dots & C_{1f}^{g\xi} & C_{11}^{g\eta} & \dots & C_{1f}^{g\eta} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ C_{f1}^{gg} & \dots & C_{ff}^{gg} & C_{f1}^{g\xi} & \dots & C_{ff}^{g\xi} & C_{f1}^{g\eta} & \dots & C_{ff}^{g\eta} \\ \\ C_{11}^{\xi g} & \dots & C_{1f}^{\xi g} & C_{11}^{\xi\xi} & \dots & C_{1f}^{\xi\xi} & C_{11}^{\xi\eta} & \dots & C_{1f}^{\xi\eta} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ C_{f1}^{\xi g} & \dots & C_{ff}^{\xi g} & C_{f1}^{\xi\xi} & \dots & C_{ff}^{\xi\xi} & C_{f1}^{\xi\eta} & \dots & C_{ff}^{\xi\eta} \\ \\ C_{11}^{\eta g} & \dots & C_{1f}^{\eta g} & C_{11}^{\eta\xi} & \dots & C_{1f}^{\eta\xi} & C_{11}^{\eta\eta} & \dots & C_{1f}^{\eta\eta} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ C_{f1}^{\eta g} & \dots & C_{ff}^{\eta g} & C_{f1}^{\eta\xi} & \dots & C_{ff}^{\eta\xi} & C_{f1}^{\eta\eta} & \dots & C_{ff}^{\eta\eta} \end{bmatrix} ; \quad (15-6)$$

if N is to be computed at m points, then C_{st} consists of m rows of form (15-5).

The quantities N , Δg , ξ , η are related to the anomalous potential T by

$$N = \frac{1}{\gamma_0} T, \quad (15-7)$$

$$\Delta g = -\frac{\partial T}{\partial r} - \frac{2}{r} T, \quad (15-8)$$

$$\xi = \frac{1}{\gamma_0 r} \frac{\partial T}{\partial \theta}, \quad \eta = -\frac{1}{\gamma_0 r \sin \theta} \frac{\partial T}{\partial \lambda}. \quad (15-9)$$

The first equation is Bruns' formula (2-31), with normal gravity γ replaced by a constant mean value γ_0 ; (15-8) and (15-9) are (11-1) and (11-2) respectively. Therefore, covariance propagation (11-14) gives

$$\text{cov}[N(P), \Delta g(Q)] = \frac{1}{\gamma_0} \left(-\frac{\partial K}{\partial r'} - \frac{2}{r'} K \right)$$

$$\text{cov}[N(P), \xi(Q)] = \frac{1}{\gamma_0^2 r'} \frac{\partial K}{\partial \theta'},$$

$$\text{cov}[N(P), \eta(Q)] = -\frac{1}{\gamma_0^2 r' \sin \theta'} \frac{\partial K}{\partial \lambda'},$$

$$\begin{aligned} \text{cov}[\Delta g(P), \Delta g(Q)] &= \left(-\frac{\partial}{\partial r} - \frac{2}{r} \right) \left(-\frac{\partial K}{\partial r'} - \frac{2}{r'} K \right) \\ &= \frac{\partial^2 K}{\partial r \partial r'} + \frac{2}{r'} \frac{\partial K}{\partial r} + \frac{2}{r} \frac{\partial K}{\partial r'} + \frac{4}{rr'} K, \end{aligned}$$

$$\begin{aligned}
\text{cov}[\Delta g(P), \xi(Q)] &= \left(-\frac{\partial}{\partial r} - \frac{2}{r} \right) \frac{1}{r'} \frac{\partial K}{\partial \theta'} \frac{1}{\gamma_0} \\
&= - \left(\frac{1}{r'} \frac{\partial^2 K}{\partial r \partial \theta'} + \frac{2}{rr'} \frac{\partial K}{\partial \theta'} \right) \frac{1}{\gamma_0}, \\
\text{cov}[\Delta g(P), \eta(Q)] &= \left(-\frac{\partial}{\partial r} - \frac{2}{r} \right) \left(-\frac{1}{r' \sin \theta'} \frac{\partial K}{\partial \lambda'} \right) \frac{1}{\gamma_0} \\
&= \left(\frac{1}{r' \sin \theta'} \frac{\partial^2 K}{\partial r \partial \lambda'} + \frac{2}{rr' \sin \theta'} \frac{\partial K}{\partial \lambda'} \right) \frac{1}{\gamma_0}, \\
\text{cov}[\xi(P), \xi(Q)] &= \frac{1}{rr'} \frac{\partial^2 K}{\partial \theta \partial \theta'} \frac{1}{\gamma_0^2}, \\
\text{cov}[\xi(P), \eta(Q)] &= -\frac{1}{rr' \sin \theta'} \frac{\partial^2 K}{\partial \theta \partial \lambda'} \frac{1}{\gamma_0^2}, \\
\text{cov}[\eta(P), \eta(Q)] &= \frac{1}{rr' \sin \theta \sin \theta'} \frac{\partial^2 K}{\partial \lambda \partial \lambda'} \frac{1}{\gamma_0^2}.
\end{aligned} \tag{15-10}$$

Here

$$K = K(P, Q) = K(r, r', \psi) \tag{15-11}$$

is the covariance function for the anomalous potential T . It is a function of the coordinates (r, θ, λ) of P and (r', θ', λ') of Q in the following manner: it depends on the radial coordinates r and r' directly, but the dependence on the angular coordinates (θ, λ) and (θ', λ') is indirect through ψ given by

$$\cos \psi = \cos \theta \cos \theta' + \sin \theta \sin \theta' \cos(\lambda' - \lambda), \tag{15-12}$$

as the expression (10-3) shows. Therefore,

$$\begin{aligned}
\frac{\partial K}{\partial \theta} &= \frac{\partial K}{\partial \psi} \frac{\partial \psi}{\partial \theta}, & \frac{\partial K}{\partial \lambda} &= \frac{\partial K}{\partial \psi} \frac{\partial \psi}{\partial \lambda}, \\
\frac{\partial K}{\partial \theta'} &= \frac{\partial K}{\partial \psi} \frac{\partial \psi}{\partial \theta'}, & \frac{\partial K}{\partial \lambda'} &= \frac{\partial K}{\partial \psi} \frac{\partial \psi}{\partial \lambda'},
\end{aligned}$$

where the partial derivatives $\partial \psi / \partial \theta, \dots, \partial \psi / \partial \lambda'$ are obtained by differentiating (15-12) according to the procedure used in deriving Vening Meinesz' formula (cf. Heiskanen and Moritz, 1967, p.113).

Now the elements of the signal covariance matrices (15-5) and (15-6) can be expressed in terms of the covariances (15-10):

$$\begin{aligned}
C_{P1}^{Ng} &= \text{cov}[N(P), \Delta g(P_1)] \quad , \\
C_{P1}^{N\xi} &= \text{cov}[N(P), \xi(P_1)] \quad , \\
C_{P1}^{N\eta} &= \text{cov}[N(P), \eta(P_1)] \quad , \\
C_{ij}^{gg} &= \text{cov}[\Delta g(P_i), \Delta g(P_j)] \quad , \\
C_{ij}^{g\xi} &= \text{cov}[\Delta g(P_i), \xi(P_j)] \quad , \\
C_{ij}^{g\eta} &= \text{cov}[\Delta g(P_i), \eta(P_j)] \quad , \\
C_{ij}^{\xi\xi} &= \text{cov}[\xi(P_i), \xi(P_j)] \quad , \\
C_{ij}^{\xi\eta} &= \text{cov}[\xi(P_i), \eta(P_j)] \quad , \\
C_{ij}^{\eta\eta} &= \text{cov}[\eta(P_i), \eta(P_j)] \quad ;
\end{aligned} \tag{15-13}$$

there is obviously

$$C_{ij}^{\xi g} = C_{ji}^{g\xi} \quad , \quad C_{ij}^{ng} = C_{ji}^{gn} \quad , \quad C_{ij}^{n\xi} = C_{ji}^{\xi n} \quad , \tag{15-14}$$

The expressions (15-13) signify that the spherical coordinates of the points P or Q in (15-10) are to be replaced by the coordinates (R, θ_i, λ_i) of P_i or (R, θ_j, λ_j) of P_j as indicated; we have $r_i = r_j = R$ since the points are considered to be situated at sea level, characterized, in the spherical approximation, by $r = R$. In the first three covariance expressions (15-13), P denotes the point at which N is to be computed; also for this point we have $r = R$.

Now we have determined all quantities necessary for evaluating the prediction formula (15-1) and the corresponding error covariance matrix (14-42). It is also easy to introduce appropriate modifications for other data configurations, for instance, for the case in which the astrogeodetic stations differ from the gravity points.

Numerical aspects of least-squares collocation will be discussed in sec. 18. The numerous literature on the problem considered in the present section includes (Moritz, 1970), (Grafarend, 1971), (Gentry and Nash, 1972), (Heitz and Tscherning, 1972), (Grafarend and Offermans, 1975), (Tscherning, 1975b), and (Lachapelle, 1975).

16. LEAST-SQUARES COLLOCATION WITH PARAMETERS

An ultimate generalization of the linear collocation models (11-4) and (14-4) is achieved by introducing functional (non-random) parameters forming a p -vector X (as usual, all vectors are column vectors unless the contrary is stated). Thus (14-4) is generalized as follows:

$$l = AX + BT + n \quad (16-1)$$

where A is a given $q \times p$ matrix expressing the effect of the parameters X on the observations l_i ; it is sometimes called "sensitivity matrix". The expression AX is usually obtained by linearizing an originally non-linear function of the p parameters; it represents the "systematic" or "parametric" part of l . The q -vector

$$t = BT \quad (16-2)$$

gives the "signal part" of l , where B again comprises the q functionals (11-5). As before, the q -vector n denotes the measuring errors (the "noise"). We assume that $q > p$ and that A has full rank.

By means of the abbreviation (16-2) we may write the observation equation (16-1) as

$$l = AX + t + n. \quad (16-3)$$

In sec.18 we shall see that all geodetic observations, after linearization, can be expressed in this form, so that the linear representation (16-3) is, in fact, quite general.

As in sec.14, the quantities t and n are centered, satisfying (14-14) and (14-15).

For $A = 0$, eq.(16-3) reduces to (14-6), which is the vectorial observation equation for least-squares collocation (prediction) without parameters; for $B = 0$ or $t = 0$ we get

$$l = AX + n, \quad (16-4)$$

which is the linear form of the observation equation in usual least-squares adjustment by parameters. We may thus say that (16-1) or (16-3) represents a model which is a synthesis between adjustment and prediction.

In contrast to least-squares adjustment, the model (16-3) contains a second term which is (formally) treated as a random quantity, namely the

signal vector t as given by (16-2). Let us estimate another centered signal vector s of m components; it has the same meaning as in the preceding sections. Both vectors s and t comprise elements of the anomalous gravitational field, which are linear functionals of the potential T ; these vectors are related by their signal covariances forming the matrices

$$C_{ss} = \text{cov}(s, s) = M\{ss^T\}, \quad (16-5)$$

$$C_{st} = \text{cov}(s, t) = M\{st^T\}, \quad (16-6)$$

$$C_{tt} = \text{cov}(t, t) = M\{tt^T\}. \quad (16-7)$$

They must be derived rigorously from one basic covariance function $K(P, Q)$ in order to preserve the analytical field structure, as explained in sec.14.

We also assume the covariance matrix of the measuring errors,

$$C_{nn} = E\{nn^T\}, \quad (16-8)$$

to be given; for an explanation of the symbols M and E cf. sec.14.

From ordinary least-squares adjustment we know that the estimates satisfy two different but equivalent minimum conditions, both of which have been given already by Gauss: least squares and minimum variance. The minimum variance condition has been used to derive the basic prediction formula (9-28). Here we shall try to use an appropriate least-squares condition.

The well-known least-squares condition for the adjustment model (16-4) is

$$n^T C_{nn}^{-1} n = \text{minimum}. \quad (16-9)$$

Let us find a suitable generalization for the present model (16-3).

All quantities treated as random may be combined into the $m + q$ vector

$$v = \begin{bmatrix} s \\ n \end{bmatrix} = [s_1 \ s_2 \ \dots \ s_m \ n_1 \ n_2 \ \dots \ n_q]^T. \quad (16-10)$$

This vector will also contain the signals t if we stipulate that $m \geq q$ and that the first q components of the vector s are identical to the components of the vector t . Thus s has the form

$$s = \begin{bmatrix} t \\ u \end{bmatrix} , \quad (16-11)$$

the vector u comprising the $m - q$ signals which are to be estimated. Usually we have $m > q$, but the following algorithm works even for $m = q$, u being absent.

The covariance matrix of the vector v has the form, in view of (16-10), of a partitioned matrix

$$C_{vv} = \begin{bmatrix} C_{ss} & 0 \\ 0 & C_{nn} \end{bmatrix} ; \quad (16-12)$$

the off-diagonal terms are zero since signal and noise are uncorrelated (p.101).

The inverse of (16-12) is

$$C_{vv}^{-1} = \begin{bmatrix} C_{ss}^{-1} & 0 \\ 0 & C_{nn}^{-1} \end{bmatrix} , \quad (16-13)$$

by well-known relations for partitioned matrices (according to our constant assumption, all matrices have full rank!).

An appropriate generalization of (16-9) is

$$v^T C_{vv}^{-1} v = \text{minimum} , \quad (16-14)$$

which by (16-10) and (16-13) becomes

$$s^T C_{ss}^{-1} s + n^T C_{nn}^{-1} n = \text{minimum} . \quad (16-15)$$

This is the *minimum principle for least-squares collocation* which we shall use for deriving optimal estimates for X and s ; in the next section we shall prove that these estimates also satisfy the second Gaussian condition, minimum variance.

The minimum problem (16-15) is to be solved with the side condition (16-3), which can easily be reformulated in terms of s as

$$l = AX + Us + n . \quad (16-16)$$

Here U is a $q \times m$ matrix partitioned into

$$U = [I \quad 0] , \quad (16-17)$$

where I is the $q \times q$ matrix and 0 is a $q \times (m-q)$ zero matrix. In view of (16-11) it is clear that

$$Us = t . \quad (16-18)$$

This minimum problem is solved using the method of Lagrange multipliers, which form a q -vector k . We are thus to find the unconditional minimum of the function

$$\phi(s, n, X) = \frac{1}{2} s^T C_{ss}^{-1} s + \frac{1}{2} n^T C_{nn}^{-1} n - k^T (AX + Us + n - t) . \quad (16-19)$$

For this purpose we form the differential

$$\begin{aligned} d\phi &= s^T C_{ss}^{-1} ds + n^T C_{nn}^{-1} dn - k^T (AdX + Uds + dn) , \\ &= (s^T C_{ss}^{-1} - k^T U) ds + (n^T C_{nn}^{-1} - k^T) dn - k^T AdX , \end{aligned} \quad (16-20)$$

which must be zero for arbitrary ds , dn , and dX . This implies

$$s^T C_{ss}^{-1} - k^T U = 0 , \quad (16-21)$$

$$n^T C_{nn}^{-1} - k^T = 0 , \quad (16-22)$$

$$k^T A = 0 . \quad (16-23)$$

The first equation gives

$$s^T = k^T U C_{ss}$$

or

$$s = C_{ss} U^T k , \quad (16-24)$$

since $C_{ss}^T = C_{ss}$ because of symmetry. From the second equation we get in the same way

$$n = C_{nn}k . \quad (16-25)$$

We write (16-16) in the form

$$Us + n = 1 - AX$$

and substitute (16-24) and (16-25), obtaining

$$(UC_{ss}U^T + C_{nn})k = 1 - AX . \quad (16-26)$$

Now, in view of (16-11),

$$C_{ss} = \begin{bmatrix} C_{tt} & C_{tu} \\ C_{ut} & C_{uu} \end{bmatrix} , \quad (16-27)$$

so that by (16-17)

$$UC_{ss}U^T = C_{tt} . \quad (16-28)$$

Let us use the abbreviation

$$\bar{C} = C_{tt} + C_{nn} ; \quad (16-29)$$

this is nothing else than the total covariance matrix of 1 , denoted by C_{11} in (14-23); it is the sum of the covariance matrices of the signal t and the noise n in (16-3).

Hence (16-26) becomes

$$\bar{C}k = 1 - AX ,$$

so that

$$k = \bar{C}^{-1}(1 - AX) . \quad (16-30)$$

This is substituted into (16-23), written as

$$A^T k = 0 ,$$

with the result

$$A^T \bar{C}^{-1} A X = A^T \bar{C}^{-1} 1$$

or

$$X = (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} 1 \quad (16-31)$$

Combining (16-24) and (16-30) we obtain

$$s = C_{ss} U^T \bar{C}^{-1} (1 - AX) \quad (16-32)$$

By (16-17) and (16-27) we have

$$C_{ss} U^T = \begin{bmatrix} C_{tt} & C_{tu} \\ C_{ut} & C_{uu} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} C_{tt} \\ C_{ut} \end{bmatrix} = C_{st} \quad (16-33)$$

since

$$C_{st} = M\{st^T\} = M\left\{\begin{bmatrix} t \\ u \end{bmatrix} t^T\right\} = \begin{bmatrix} M\{tt^T\} \\ M\{ut^T\} \end{bmatrix} = \begin{bmatrix} C_{tt} \\ C_{ut} \end{bmatrix} \quad (16-34)$$

Thus (16-32) becomes

$$s = C_{st} \bar{C}^{-1} (1 - AX) \quad (16-35)$$

Since (16-31) and (16-35) give estimates for X and s , we write these equations with our usual notation for estimates:

$$\hat{X} = (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} 1 \quad (16-36)$$

$$\hat{s} = C_{st} \bar{C}^{-1} (1 - A\hat{X}) \quad (16-37)$$

where \bar{C} is the total covariance matrix (16-29).

These equations solve our problem: first the estimated values of the parameters X are computed from (16-36), then the estimated (predicted and/or filtered) values of the signal s are obtained from (16-37).

Equation (16-36) is completely analogous to the equation determining the parameters in least-squares adjustment, with the important difference that in adjustment by parameters we have the noise covariance matrix C_{nn} instead of C , whereas in collocation the signal covariance matrix C_{tt} enters as well.

Equation (16-37) is analogous to the prediction equation (9-28) and reduces to (14-27) if there are no systematic parameters.

Thus we see again that the present model of least-squares collocation with parameters combines least-squares adjustment and least-squares prediction into a unified scheme.

Relation to least-squares adjustment. Formally the present mathematical model can be reduced to a least-squares adjustment by condition equations with parameters. In fact, using (16-10), we can write (16-16) in the form

$$l = AX + Vv, \quad (16-38)$$

where

$$V = [U \quad I], \quad (16-39)$$

I being the $q \times q$ unit matrix. The problem of solving (16-38) under the minimum condition (16-14), written as

$$v^T P v = \text{minimum}, \quad (16-40)$$

with $P = C_{vv}^{-1}$, is clearly formally identical to the problem of adjustment of condition equations with parameters, also called "general case of least-squares adjustment" (Wolf, 1968, p.133).

Still, it differs from an adjustment problem in the strict sense (provided $m > q$). In adjustment computations, the conditions connect all observations and, therefore, all relevant random quantities: all residuals v . The principle $v^T P v = \text{minimum}$ contains only those residuals which also occur explicitly in the condition equations.

In the present problem, however, the quantities that are most important here, namely the signals to be predicted, do not enter at all into the observation equations (16-3). This seems to be contradicted by (16-38), but the contradiction is only apparent. Consider (16-11), in which t com-

prises the signal parts of the observations t , and u consists of the signals to be predicted. If the vector v is partitioned into

$$v = \begin{bmatrix} t \\ u \\ n \end{bmatrix}, \quad (16-41)$$

combining (16-10) and (16-11), then the corresponding partitioning of v is

$$V = [I \quad 0 \quad I], \quad (16-42)$$

by (16-17) and (16-39). Thus in (16-38) we have

$$Vv = t + n; \quad (16-43)$$

the vector u does not enter into the observation equation, although it occurs in the minimum principle (16-40)!

Our problem thus contains additional random variables u which are related to the observations only through the joint covariances, which is characteristic for prediction (sec.9). Formally this makes no difference in (16-38); it only means that the matrix V contains some all-zero columns, but conceptually it is quite significant: we have a combined problem of adjustment and prediction.

This is true at least if $m > q$ that is, as long as we wish to predict other signals besides the "signal parts" of the observations. As we shall see in sec.18, this is the case relevant to the determination of the gravity field. In the limiting case $m = q$ we have $s = t$, only the signal parts of the observation are computed by removing systematic effects AX and the noise n as much as possible. This is the case of "pure filtering"; this limiting case, indeed, formally reduces to an adjustment by condition equations with parameters.

A genuine and complete reduction of the general model of least-squares collocation to an adjustment problem is only possible in infinite-dimensional Hilbert space, as we shall see in sec.29.

Properties of the solution. The solution expressed by (16-36) and (16-37) has the following properties:

- (a) The result is independent of the number m of signal quantities s to be estimated.

- (b) Both observed and estimated quantities can be heterogeneous, provided all required covariances are known.
- (c) The method is invariant with respect to linear transformations of the data or of the results.
- (d) The solution is optimal in the sense that it gives the most accurate results obtainable on the basis of the given data.

As for (a), the vector u of predicted signal may consist of one component, or it may consist of a thousand components. The result for the same signal quantity will always be the same, for the following reason. Eq. (16-36) depends on the observations only; the predicted signals u do not enter at all. In (16-37), the term $\bar{C}^{-1}(1-A\bar{X})$ likewise depends only on the observations; each component of the vector s is obtained individually since only the k -th row of the matrix C_{st} affects the k -th component of s .

Thus it suffices to include, in the vector s and in the minimum condition (16-15), only as many signals u as we wish to compute. This procedure looks somewhat arbitrary. In fact, (16-15) is only a "shortened version" of a more complete minimum principle in Hilbert space, as we shall see in sec.30. The present elementary treatment of collocation is intended to use elementary matrix calculus as much as possible, avoiding Hilbert space (which only enters indirectly through covariance propagation).

As for (b), the (physical or mathematical) nature of the quantities s and n is irrelevant. All we have to require is that they are centered (having average zero) and that all covariance matrices are known.

Property (b) follows from the corresponding invariance properties of least-squares adjustment, for (16-36), and least-squares prediction, for (16-37); cf. sec.12.

Proof that (16-15) is satisfied. We have derived the present solution from the minimum principle (16-15) by putting the differential $d\phi$, given by (16-20), equal to zero. This, however, is only a necessary but not sufficient condition for a minimum. Let us, therefore, prove that we have indeed a minimum of (16-15).

Consider the least-squares estimates \hat{X} and \hat{s} , as given by (16-36) and (16-37). There is also a least-squares estimate \hat{n} for the noise, given by

$$\hat{n} = C_{nn}\bar{C}^{-1}(1 - A\hat{X}) ; \quad (16-44)$$

this is a direct consequence of (16-25) and (16-30). Consider the vector

$$\hat{v} = \begin{bmatrix} \hat{s} \\ \hat{h} \end{bmatrix} ; \quad (16-45)$$

clearly \hat{X} and \hat{v} satisfy (16-38) :

$$A\hat{X} + V\hat{v} = 1 . \quad (16-46)$$

Besides the least-squares estimates \hat{X} and \hat{v} consider arbitrary other estimates \bar{X} and \bar{v} which also satisfy this condition:

$$A\bar{X} + V\bar{v} = 1 . \quad (16-47)$$

Put

$$\begin{aligned} \bar{X} &= \hat{X} + X_1 , \\ \bar{v} &= \hat{v} + v_1 . \end{aligned} \quad (16-48)$$

By subtracting (16-46) and (16-47) we see that

$$AX_1 + Vv_1 = 0 . \quad (16-49)$$

We then have, with $P = C_{vv}^{-1}$:

$$\bar{v}^T P \bar{v} = (\hat{v}^T + v_1^T) P (\hat{v} + v_1) = \hat{v}^T P \hat{v} + v_1^T P \hat{v} + \hat{v}^T P v_1 + v_1^T P v_1 . \quad (16-50)$$

Let us prove that

$$v_1^T P \hat{v} = 0 = \hat{v}^T P v_1 . \quad (16-51)$$

In fact, the estimates (16-36) and (16-44) may be combined as

$$\hat{v} = C_{vv} V^T \bar{C}^{-1} (1 - A\hat{X}) , \quad (16-52)$$

since by (16-12) and (16-39) ,

$$C_{vv}V^T = \begin{bmatrix} C_{ss} & 0 \\ 0 & C_{nn} \end{bmatrix} \begin{bmatrix} U^T \\ I \end{bmatrix} = \begin{bmatrix} C_{ss}U^T \\ C_{nn} \end{bmatrix} = \begin{bmatrix} C_{st} \\ C_{nn} \end{bmatrix} \quad (16-53)$$

by (16-33). By (16-52) we get

$$v_1^T P \hat{v} = v_1^T C_{vv}^{-1} \hat{v} = v_1^T V^T \bar{C}^{-1} (1 - A\hat{X}) .$$

From (16-49) there follows

$$v_1^T V^T = -X_1^T A^T , \quad (16-54)$$

whence, together with (16-36),

$$\begin{aligned} v_1^T P \hat{v} &= -X_1^T A^T \bar{C}^{-1} (1 - A\hat{X}) \\ &= -X_1^T A^T \bar{C}^{-1} (1 - A(A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} 1) \\ &= -X_1^T (A^T \bar{C}^{-1} 1 - A^T \bar{C}^{-1} A (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} 1) = 0 , \end{aligned}$$

which proves (16-51).

Hence (16-50) reduces to

$$\bar{v}^T P \bar{v} = \hat{v}^T P \hat{v} + v_1^T P v_1 . \quad (16-55)$$

From the positive definiteness of C_{vv} it follows that also $P = C_{nn}^{-1}$ is positive-definite; hence, by (9-26),

$$v_1^T P v_1 \geq 0 .$$

Thus

$$\bar{v}^T P \bar{v} \geq \hat{v}^T P \hat{v} ,$$

which shows that the least-squares solution gives indeed a minimum of $v^T P v$,

17. ACCURACY

In this section we shall first derive expressions for the standard errors and error covariances of the quantities X and s obtained by an arbitrary linear unbiased estimation, then we shall specialize these expressions for least-squares collocation, and finally we shall show that collocation gives indeed optimum estimates in the sense that in this case the standard errors are the smallest that are possible for any linear estimation method.

Linear estimates. Consider any linear estimates for X and s , that is, expressions of the form

$$\hat{s} = L\hat{l} + a, \quad (17-1)$$

$$\hat{X} = G\hat{l} + b, \quad (17-2)$$

where L is a $m \times q$ matrix, G is a $p \times q$ matrix, a is a m -vector, and b is a p -vector. The quantities L , G , a , b are assumed to be independent of \hat{l} , so that (17-1) and (17-2) represent the estimated values of s and X as linear functions of the measurements \hat{l} . This is the meaning of the term "linear estimate".

These estimates must reasonably satisfy the same relations as the original "true" values. In the sequel we shall always denote true values by an overbar.

Thus (16-3) may be written for true values as

$$\bar{l} = A\bar{X} + \bar{z}, \quad (17-3)$$

if we put

$$\bar{z} = \bar{t} + \bar{n}; \quad (17-4)$$

for their estimates we have the analogous relations

$$\hat{l} = A\hat{X} + \hat{z}, \quad (17-5)$$

$$\hat{z} = \hat{t} + \hat{n}. \quad (17-6)$$

Let us now postulate that the estimates (17-1) and (17-2) be *unbiased*. That is, in terms of the average $\bar{}$ introduced in sec. 14 we must have

$$\bar{E}(\xi) = 0 , \quad (17-7)$$

$$\bar{E}(\hat{X}) = X , \quad (17-8)$$

where X is the true value of the parameter vector. This latter equation is the usual condition for unbiasedness for parameter estimation, whereas (17-7) represents the natural generalization for the estimation of a centered random quantity, already used in sec.9.

Substitute (17-3) into (17-1) to obtain

$$\xi = LAX + Lz + a ,$$

and form the average \bar{E} . The result is

$$\bar{E}(\xi) = LAX + L\bar{E}(z) + a ,$$

since X and a are nonrandom quantities. By (17-7), (14-14), and (14-15) this reduces to

$$0 = LAX + a . \quad (17-9)$$

For the estimation formula (17-1) to be meaningful, the vector a must be given beforehand and cannot depend on the true value X , which is forever unknown. Eq. (17-9) will be satisfied for arbitrary X if and only if

$$LA = 0 , \quad a = 0 . \quad (17-10)$$

Let us now investigate (17-2) in a similar way. Substitute (17-3) into (17-2), obtaining

$$\hat{X} = GAX + Gz + b ,$$

and form the average \bar{E} , using (17-8). The result is

$$X = GAX + b$$

or

$$(I - GA)X - b = 0 ,$$

where I denotes the unit matrix. Reasoning as before, we obtain the conditions

$$GA = I, \quad b = 0, \quad (17-11)$$

which are likewise necessary and sufficient.

The unbiased linear estimates are thus

$$\hat{s} = L\mathbf{l} \quad \text{with} \quad LA = 0, \quad (17-12)$$

$$\hat{X} = G\mathbf{l} \quad \text{with} \quad GA = I. \quad (17-13)$$

Let us now put

$$L = H(I - AG), \quad (17-14)$$

for which the condition $LA = 0$ is automatically satisfied:

$$LA = H(I - AG)A = H(A - AGA) = H(A - A) = 0,$$

since $GA = I$ by (17-11); it may be shown that all matrices L satisfying $LA = 0$ can be represented in the form (17-14) (take, e.g., $H = L$).

Then (17-12) becomes

$$\hat{s} = L\mathbf{l} = H(I - AG)\mathbf{l} = H(\mathbf{l} - AG\mathbf{l}) = H(\mathbf{l} - A\hat{X})$$

by (17-13).

In this way we finally obtain the general expression for unbiased linear estimates:

$$\hat{X} = G\mathbf{l}, \quad (17-15)$$

$$\hat{s} = H(\mathbf{l} - A\hat{X}), \quad (17-16)$$

where the $m \times q$ matrix H is arbitrary and the $p \times q$ matrix G satisfies the condition

$$GA = I. \quad (17-17)$$

Errors of estimation. The individual error of the estimates is defined as the difference: estimated value minus true value. Thus the individual error of \hat{X} is

$$\epsilon_X = \hat{X} - X = G\mathbf{l} - X = GAX + Gz - X = Gz$$

by (17-3) and (17-17), and the individual error of \mathbf{s} is

$$\epsilon_s = \mathbf{\hat{s}} - \mathbf{s} = \mathbf{L}\mathbf{l} - \mathbf{s} = \mathbf{L}\mathbf{A}\mathbf{x} + \mathbf{L}\mathbf{z} - \mathbf{s} = \mathbf{L}\mathbf{z} - \mathbf{s}$$

because of (17-10). Summarizing we have

$$\epsilon_X = \mathbf{G}\mathbf{z}, \quad (17-18)$$

$$\epsilon_s = \mathbf{L}\mathbf{z} - \mathbf{s}. \quad (17-19)$$

Let us now pass on to the corresponding standard errors and error covariances. We compute (T denotes the transpose as usual)

$$\epsilon_X \epsilon_X^T = \mathbf{G}\mathbf{z}\mathbf{z}^T\mathbf{G}^T \quad (17-20)$$

$$\begin{aligned} \epsilon_s \epsilon_s^T &= (\mathbf{L}\mathbf{z} - \mathbf{s})(\mathbf{z}^T\mathbf{L}^T - \mathbf{s}^T) \\ &= \mathbf{s}\mathbf{s}^T - \mathbf{L}\mathbf{z}\mathbf{s}^T - \mathbf{s}\mathbf{z}^T\mathbf{L}^T + \mathbf{L}\mathbf{z}\mathbf{z}^T\mathbf{L}^T. \end{aligned} \quad (17-21)$$

Let us now form the average \bar{E} of these two equations. The matrices

$$\mathbf{E}_{XX} = \bar{E}\{\epsilon_X \epsilon_X^T\} \quad (17-22)$$

$$\mathbf{E}_{ss} = \bar{E}\{\epsilon_s \epsilon_s^T\} \quad (17-23)$$

are the error covariance matrices of the vectors \mathbf{X} and \mathbf{s} (the "E" in \mathbf{E}_{XX} and \mathbf{E}_{ss} comes from Error covariance matrix, whereas in $\bar{E}(\cdot)$ it comes from Expectation!). We further have

$$\bar{E}\{\mathbf{s}\mathbf{s}^T\} = \mathbf{C}_{ss} = \mathbf{M}\{\mathbf{s}\mathbf{s}^T\} \quad (17-24)$$

by (14-17), and

$$\bar{E}\{\mathbf{z}\mathbf{z}^T\} = \mathbf{C}_{tt} + \mathbf{C}_{nn} = \bar{\mathbf{C}} \quad (17-25)$$

by (16-29), since

$$\mathbf{z} = \mathbf{t} + \mathbf{n} = \mathbf{I} - \mathbf{A}\mathbf{X} \quad (17-26)$$

is the "random part" of \mathbf{I} . Similarly,

$$E\{zs^T\} = C_{zs} = C_{ts} \quad (17-27)$$

because signal and noise are uncorrelated so that $C_{ns} = 0$. We also put, as usual,

$$C_{st} = C_{ts}^T. \quad (17-28)$$

With these notations, the averages \bar{E} of (17-20) and (17-21) are

$$E_{xx} = G\bar{C}G^T, \quad (17-29)$$

$$E_{ss} = C_{ss} - LC_{ts} - C_{st}L^T + L\bar{C}L^T. \quad (17-30)$$

These expressions give the error covariance matrices for any linear unbiased estimation of x and s . The diagonal terms represent the error variances (the squares of the standard errors) of the estimated quantities; the off-diagonal terms represent the error covariances.

We may also consider the cross-covariance matrices

$$E_{xs} = E\{\epsilon_x \epsilon_s^T\}, \quad (17-31)$$

$$E_{sx} = E\{\epsilon_s \epsilon_x^T\} = E_{xs}^T. \quad (17-32)$$

By (17-18) and (17-19) we have

$$\epsilon_x \epsilon_s^T = -Gzs^T + Gzz^T L^T$$

and hence

$$E_{xs} = -GC_{ts} + G\bar{C}L^T = E_{sx}^T. \quad (17-33)$$

Least-squares collocation. Let us now specialize these general expressions to the case of least-squares collocation. Comparing (17-15) and (17-16) to (16-36) and (16-37) we see that here

$$G = (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1}, \quad (17-34)$$

$$H = C_{st} \bar{C}^{-1} .$$

(17-35)

The condition (17-17) for an unbiased estimate is clearly satisfied. By means of (17-34), the expression (17-29) becomes

$$\begin{aligned} E_{xx} &= G \bar{C} G^T = \\ &= (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} \bar{C} \bar{C}^{-1} A (A^T \bar{C}^{-1} A)^{-1} \\ &= (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} A (A^T \bar{C}^{-1} A)^{-1} \\ &= (A^T \bar{C}^{-1} A)^{-1} . \end{aligned} \quad (17-36)$$

By (17-14) with (17-34) and (17-35) we have

$$L = C_{st} \bar{C}^{-1} \left[I - A (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} \right] , \quad (17-37)$$

so that

$$L C_{ts} = C_{st} \bar{C}^{-1} \left[I - A (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} \right] C_{ts} , \quad (17-38)$$

It is readily seen that

$$C_{st} L^T = L C_{ts} . \quad (17-39)$$

By (17-37) we get

$$\begin{aligned} L \bar{C} L^T &= C_{st} \bar{C}^{-1} \left[I - A (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} \right] \bar{C} \left[I - \bar{C}^{-1} A (A^T \bar{C}^{-1} A)^{-1} A^T \right] \bar{C}^{-1} C_{ts} \\ &= C_{st} \bar{C}^{-1} \left[I - A (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} \right]^2 C_{ts} . \end{aligned}$$

By direct computation one immediately verifies that the matrix expression between brackets is idempotent, that is,

$$\left[I - A (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} \right]^2 = I - A (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} . \quad (17-40)$$

Thus

$$L\bar{C}L^T = C_{st}\bar{C}^{-1}\left[I - A(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}\right]C_{ts} , \quad (17-41)$$

so that all three expressions (17-38), (17-39) and (17-41) entering into (17-30) are equal. Therefore, (17-30) reduces to

$$E_{ss} = C_{ss} - LC_{ts} . \quad (17-42)$$

Finally, (17-33) takes the form

$$\begin{aligned} E_{xs} &= -GC_{ts} + G\bar{C}L^T = \\ &= -(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}C_{ts} + (A^T\bar{C}^{-1}A)^{-1}A^TL^T = \\ &= (A^T\bar{C}^{-1}A)^{-1}A^T\left[-\bar{C}^{-1}C_{ts} + \bar{C}^{-1}C_{ts} - \bar{C}^{-1}A(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}C_{ts}\right] \\ &= -(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}A(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}C_{ts} \\ &= -(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}C_{ts} . \end{aligned} \quad (17-43)$$

Thus we have

$$E_{xx} = (A^T\bar{C}^{-1}A)^{-1} , \quad (17-44)$$

$$E_{ss} = C_{ss} - C_{st}\bar{C}^{-1}\left[I - A(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}\right]C_{ts} , \quad (17-45)$$

$$E_{xs} = -(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}C_{ts} ; \quad (17-46)$$

here we have used (17-37).

On writing (17-45) as

$$E_{ss} = C_{ss} - C_{st}\bar{C}^{-1}C_{ts} + HAE_{xx}A^TH^T , \quad (17-47)$$

where H is given by (17-35) we see that the term $HAE_{xx}A^TH^T$ represents the effect of inaccurate estimation of the parameters X . If there are no parameters, this term is zero.

similarly, (17-46) may be written more simply as

$$E_{x_B} = -E_{xx} A^T H^T . \quad (17-48)$$

The equations (16-36), (16-37), (17-44), (17-47), and (17-48) constitute the basic computational formulas for least-squares collocation, giving the estimates together with their accuracy.

Least-squares collocation as the most accurate estimation method. Let us now compare least-squares collocation to an arbitrary linear unbiased estimation method. Let the latter be characterized by the matrices \bar{G} and \bar{H} with

$$\bar{G}A = I \quad (17-49)$$

by (17-17), whereas G and H denote the corresponding least-squares matrices (17-34) and (17-35). In agreement with (17-14) we form the matrix

$$\bar{L} = \bar{H}(I - A\bar{G}) , \quad (17-50)$$

whereas L is to be given by (17-37); obviously

$$\bar{L}A = 0 . \quad (17-51)$$

Let us put

$$\bar{G} = G + g , \quad (17-52)$$

$$\bar{L} = L + \ell \quad (17-53)$$

Then

$$gA = (\bar{G} - G)A = \bar{G}A - GA = I - I = 0 , \quad (17-54)$$

$$\ell A = (\bar{L} - L)A = \bar{L}A - LA = 0 . \quad (17-55)$$

By (17-29) and (17-30), the error covariance matrices for the arbitrary estimate are given by

$$\begin{aligned}
 \bar{E}_{xx} &= \bar{G}\bar{C}\bar{G}^T = (G+g)\bar{C}(G^T+g^T) \\
 &= G\bar{C}G^T + g\bar{C}G^T + G\bar{C}g^T + g\bar{C}g^T, \\
 \bar{E}_{ss} &= C_{ss} - LC_{ts} - C_{st}L^T + L\bar{C}L^T \\
 &= C_{ss} - (L+l)C_{ts} - C_{st}(L^T+l^T) + (L+l)\bar{C}(L^T+l^T) \\
 &= C_{ss} - LC_{ts} - C_{st}L^T + L\bar{C}L^T - \\
 &\quad - lC_{ts} - C_{st}l^T + L\bar{C}l^T + l\bar{C}L^T + l\bar{C}l^T
 \end{aligned}$$

or

$$\bar{E}_{xx} = E_{xx} + g\bar{C}G^T + G\bar{C}g^T + g\bar{C}g^T, \quad (17-56)$$

$$\bar{E}_{ss} = E_{ss} + (-lC_{ts} + l\bar{C}L^T) + (-C_{st}l^T + L\bar{C}l^T) + l\bar{C}l^T, \quad (17-57)$$

where E_{xx} and E_{ss} denote the error covariance matrices for least-squares collocation.

Now, by (17-34),

$$\begin{aligned}
 G\bar{C}g^T &= (A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}\bar{C}g^T \\
 &= (A^T\bar{C}^{-1}A)^{-1}A^Tg^T = 0
 \end{aligned} \quad (17-58)$$

because

$$A^Tg^T = 0$$

as the transpose of (17-54). Similarly

$$g\bar{C}G^T = 0 \quad (17-59)$$

as the transpose of (17-58).

By (17-37) we have

$$L\bar{C}l^T = C_{st}\bar{C}^{-1}\left[I - A(A^T\bar{C}^{-1}A)^{-1}A^T\bar{C}^{-1}\right]\bar{C}l^T$$

$$\begin{aligned}
 &= C_{st} \ell^T - C_{st} \bar{C}^{-1} A (A^T \bar{C}^{-1} A)^{-1} A^T \ell^T \\
 &= C_{st} \ell^T
 \end{aligned}$$

since

$$A^T \ell^T = 0$$

as the transpose of (17-55). Thus

$$-C_{st} \ell^T + L \bar{C} \ell^T = 0, \quad (17-60)$$

and also for its transpose,

$$-\ell C_{ts} + \ell \bar{C} L^T = 0. \quad (17-61)$$

On taking these relations into account, the expressions (17-56) and (17-57) reduce to

$$\bar{E}_{XX} = E_{XX} + g \bar{C} g^T, \quad (17-62)$$

$$\bar{E}_{SS} = E_{SS} + \ell \bar{C} \ell^T. \quad (17-63)$$

To these expressions we now apply the same reasoning as in sec.9. The error variances (squares of standard errors) of the estimated quantities are the diagonal terms of the matrices E_{XX} and E_{SS} . Let us consider the r -th component of X ; its error variance is given by

$$\bar{m}_r^2 = m_r^2 + \gamma_r \bar{C} \gamma_r^T, \quad (17-64)$$

where γ_r is the r -th row of the matrix g . Since \bar{C} is positive definite, as all covariance matrices are, we have

$$\gamma_r \bar{C} \gamma_r^T \geq 0;$$

thus

$$\bar{m}_r^2 \geq m_r^2, \quad (17-65)$$

so that least-squares collocation, to which m_r corresponds, gives indeed the smallest possible standard error of any component of the vector x .

If we apply the same reasoning, word by word, to any diagonal term of the matrix E_{ss} , we find that least-squares collocation gives also the smallest possible standard error of any component of the vector s .

Thus we have proved that least-squares collocation is optimal in the sense that it gives the most accurate results that are obtainable on the basis of the available data.

This property is well known from least-squares adjustment and least-squares prediction (sec. 9); we could, of course, also have used it as a basis for deriving the solution expressed by (16-36) and (16-37).

18. APPLICATION TO PHYSICAL GEODESY

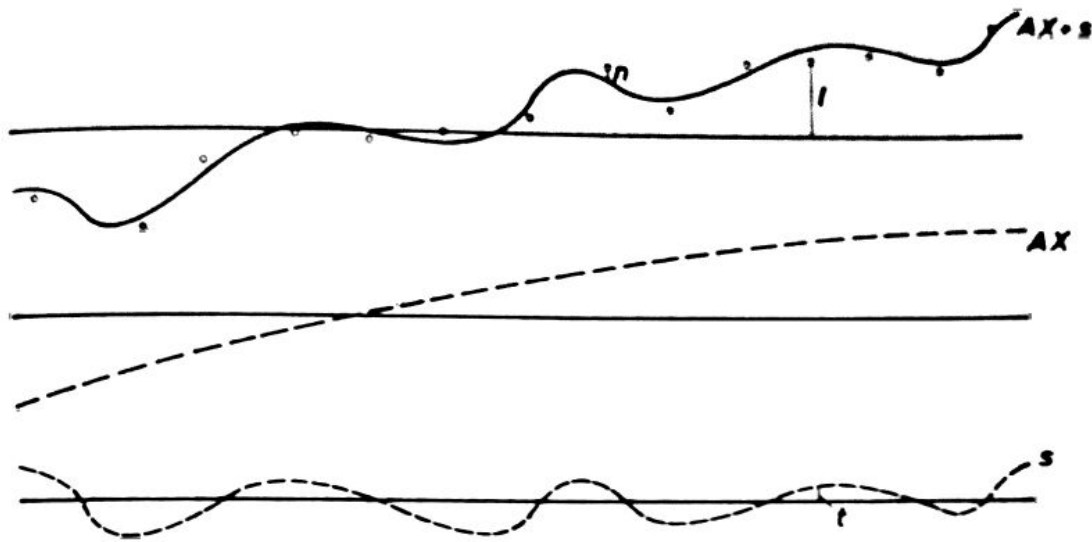
The mathematical technique presented in sections 16 and 17, least-squares collocation with parameters, can be applied in several fields. Throughout the present book we shall limit ourselves to applications to physical geodesy; for other applications cf. (Moritz, 1973a, secs. 4 and 5) and (Mikhail, 1976, chapter 14), and (Monti and Sansò, 1977).

Some particular applications to problems of physical geodesy have already been given as illustrations before, especially in sections 13 and 15; and generally, least-squares collocation, beginning with the simplest case disregarding measuring errors and systematic effects, was introduced in sec. 11 with a view to studying the gravitational field. The purpose of the present section is thus to supplement and generalize what has been said before, and to show the relevance of the general least-squares model of section 16 for the determination of the figure of the earth and of its gravitational field.

Meaning of the model. The basic mathematical model (16-3),

$$l = AX + t + n, \quad (18-1)$$

may be visualized by means of Fig. 18.1. The term AX represents a simple, regular and slowly varying curve; it is a (linear or linearized) function of a number of parameters X , for instance a polynomial whose coefficients form the vector X . Another function s , the "signal", irregularly oscillating about zero, is superimposed, giving the function $AX + s$. The problem is to determine this curve $AX + s$, shown on top (full line) by means of discrete observations l (small circles), which are further-

FIGURE 18.1. *The basic model.*

more affected by observational errors n . Denoting the signal s at the observation points by t , we arrive at the observation equation (18-1). The curve $AX + s$ to be interpolated thus consists of a "systematic part" AX , representing the general trend of the phenomenon, and a "random part" s , representing continuous irregular fluctuations, also of the physical phenomenon. In contrast to the signal s , the other random quantity n , the observational error, occurs only at the observation points and is thus discrete¹.

If we consider the determination of the parameter X as adjustment, the removal of the noise as filtering, and the computation of s at points other than the measuring points as prediction, we may say that the present model combines adjustment, filtering and prediction.

The relevance of this model for geodetic problems is made evident by mentioning some conceivable applications from quite different fields of geodesy.

1. *Gravity measurements.* Here l is the gravimeter reading, s represents the gravity anomaly Δg , n is the random measuring error, and X

¹ In certain cases, for instance in the continuous recording of sea gravity measurements, also the measuring noise may be continuous; cf. (Moritz, 1969a, sec.5). The treatment is quite analogous; here we shall limit ourselves to the practically much more important case of discrete observations.

represents systematic parameters of two different kinds: (a) the parameters of the normal gravity formula and (b) instrumental constants and other systematic effects on the measurement such as drift.

2. *Satellite observations.* Here l comprises optical or electronical measurements to artificial satellites, AX represents the normal orbit (after linearization with respect to the parameters), s represents gravitational perturbations of the orbit, and n comprises other random effects, in particular measuring errors.

3. *Transformations in geodesy and photogrammetry.* Consider two overlapping local geodetic coordinate systems. If one system is transformed into the other, there may remain residual discrepancies or distortions which are irregular but strongly correlated. Thus AX represents the transformation formula, s comprises the residual distortions, and n is the effect of measuring errors on position. This is a combined transformation and interpolation problem, which is the two-dimensional analogue to the one-dimensional problem shown in Fig.18.1. Transformation problems of precisely the same nature are frequent in photogrammetry.

4. *Graduation errors of theodolite circles.* Here l is the circle reading, AX represents the "regular" graduation error, s represents the "irregular" graduation error, and n is the reading error.

In all these problems, the measurements that make up the vector l are all of the same kind; we essentially have problems of *interpolation*. This restriction to homogeneous measurements is not essential; the extension to heterogeneous measurements which are linear functionals of a certain basic signal function is straightforward and we are thus led to *collocation* in the proper sense.

The case of physical geodesy. This extension, where the "measuring signal" t consists of linear functionals of the anomalous potential T ,

$$t = BT, \quad (18-2)$$

is basic for the determination of the gravitational field; cf. (14-2) or (16-2).

The essential fact in the present case is that the model (18-1) represents the general *observation equation for physical geodesy*. In fact, any geodetic measurement, without exception, may (after linearization) be split up into three components:

1. a systematic part AX comprising effects of the ellipsoidal reference system, station coordinates and other geometric parameters, as well as systematic measuring errors, drift, etc.;

2. a random signal part t expressing the effect of the anomalous gravity field on the measured quantity; and

3. random measuring errors n .

This corresponds to (18-1).

A general treatment of this problem will be found in sec.27; here we shall content ourselves with elementary considerations showing the basic concept. We have claimed that every geodetic observation may be represented in the form (18-1), and we shall demonstrate this by analyzing various measurable quantities l in this way. It is obviously sufficient to effect a decomposition

$$l = AX + t, \quad (18-3)$$

disregarding the measuring error n , because n can always be added afterwards.

Consider first the classical measurements of physical geodesy: magnitude and direction of the gravity vector, the first being gravity g , the second being the direction of the plumb line as defined by astronomical latitude, ϕ , and astronomical longitude, λ . We can write

$$g = \gamma + \Delta g, \quad (18-4)$$

where γ is normal gravity and Δg is the gravity anomaly. Normal gravity depends on the four parameters (denoted by p_1, p_2, p_3, p_4) defining the reference system used:

$$\gamma = \gamma(p_1, p_2, p_3, p_4), \quad (18-5)$$

to get a linear expression, introduce approximate values p_i^0 , set $p_i = p_i^0 + \delta p_i$ and linearize; then X is the vector

$$X = [\delta p_1 \quad \delta p_2 \quad \delta p_3 \quad \delta p_4]^T. \quad (18-6)$$

(In the Geodetic Reference System 1967 (IAG, 1970) we have $p_1 = a$, the semimajor axis, $p_2 = GM$, the product of gravitational constant and mass of the earth, $p_3 = J_2$, the zonal harmonic coefficient of degree 2, and $p_4 = \omega$, the rotational velocity of the earth.)

Thus in (18-4), γ represents AX , and Δg represents s .

In a similar way we may decompose the astronomical coordinates ϕ and λ :

$$\phi = \phi + \xi, \quad (18-7)$$

$$\lambda = \lambda + n \sec \phi,$$

where ϕ and λ are the corresponding geodetic coordinates which depend on the reference ellipsoid, such that

$$\phi = \phi(p_1, p_2, p_3, p_4), \quad \lambda = \lambda(p_1, p_2, p_3, p_4), \quad (18-8)$$

analogous to (18-5), and where the deflections of the vertical express the effect of the anomalous gravity field. That is, ϕ and λ constitute the systematic part AX, and ξ and $n \sec \phi$ represent the signal part s.

If we consider the parameters of the reference ellipsoid as known, then the vector (18-6) is zero, and we are able to regard Δg , ξ , n directly as observations; this case has been treated in sec. 15.

As we have seen in sec. 10, the signal field should not contain spherical harmonics of degrees 0 and 1. This implies that the reference ellipsoid is in an absolute (i.e., geocentric) position (Heiskanen and Moritz, 1967, pp.99-100).

Astrogeodetic deflections of the vertical correspond to a non-geocentric reference ellipsoid; they must therefore be transformed by shifting the ellipsoid into an absolute position (*ibid.*, p.209), before applying them in combination procedures such as the one described in sec.15.

It is, however, also possible to determine the shift parameters simultaneously by collocation: we write the astrogeodetic deflections ξ^a , n^a in the form (*ibid.*, p.213):

$$\xi^a = \frac{1}{R}(\delta x_0 \sin \phi \cos \lambda + \delta y_0 \sin \phi \sin \lambda - \delta z_0 \cos \phi) + \xi, \quad (18-9)$$

$$n^a = \frac{1}{R}(\delta x_0 \sin \lambda - \delta y_0 \cos \lambda) + n,$$

where δx_0 , δy_0 , δz_0 are the components of the shift of the reference ellipsoid and R is a mean radius of the earth.

Obviously this again fits into the model (18-3): ξ^a and n^a are observations representing l , the first terms on the right-hand side represent AX with

$$X = [\delta x_0 \quad \delta y_0 \quad \delta z_0]^T, \quad (18-10)$$

and the geocentric deflections ξ and η form the signal s .

The present method thus makes it possible to obtain at the same time:

- (a) a combined gravimetric-astrogeodetic geoid and
- (b) the shift of the astrogeodetic reference ellipsoid to its absolute position.

It may be regarded as a combination of the method described in sec.15--with respect to (a)--and the determination of the shift parameters by combining astrogeodetic and gravimetric data (Heiskanen and Moritz, 1967, sec.5-10)--with respect to (b); of course, a good shift determination (b) presupposes again a reasonable global gravity coverage.

But also any other observational quantities, which at first sight seem to be purely geometric, fit into the general scheme (18-3), for instance, measured azimuths, horizontal angles, and zenith distances.

By eqs. (5-10) and (5-13) of *ibid.*, p.186 we have, after a slight change of notation,

$$\alpha = \alpha' + \eta \tan \phi + (\xi \sin \alpha - \eta \cos \alpha) \cot \zeta, \quad (18-11)$$

where α denotes the measured ("astronomical") azimuth and α' denotes the ellipsoidal ("geodetic") azimuth; ζ is the zenith distance.

A measured horizontal angle may be considered as the difference between two azimuths:

$$\omega = \alpha_2 - \alpha_1$$

so that we have

$$\begin{aligned} \omega = \omega' + \xi (\sin \alpha_2 \cot \zeta_2 - \sin \alpha_1 \cot \zeta_1) + \\ + \eta (-\cos \alpha_2 \cot \zeta_2 + \cos \alpha_1 \cot \zeta_1), \end{aligned} \quad (18-12)$$

where ω' is the ellipsoidal horizontal angle, that is, the horizontal angle reduced to the reference ellipsoid, α_1 and ζ_1 are azimuth and zenith distance to target 1, and α_2 and ζ_2 are the same quantities for target 2.

Similarly we have for a measured zenith distance ζ (Heiskanen and Moritz, 1967, p.190):

$$\zeta = \zeta' - \xi \cos \alpha - \eta \sin \alpha, \quad (18-13)$$

where ζ' is the zenith distance reduced to the ellipsoid.

In these expressions, the ellipsoidal quantities α' , ω' , ζ' represent the "systematic" part of the observations, pertaining to the reference ellipsoid. If we vary the ellipsoidal parameters p_i by δp_i , these quantities take indeed the form AX with (18-6). The terms containing as factors the deflections of the vertical, ξ and η , represent the signal s , that is, the effect of the anomalous gravity field on the quantities under consideration.

This might be symbolized as

$$\alpha = \alpha' + s_\alpha, \quad (18-14)$$

with the "signal part of α " given by

$$s_\alpha = \eta \tan \phi + (\xi \sin \alpha - \eta \cos \alpha) \cot \zeta; \quad (18-15)$$

the quantities s_ω and s_ζ , the signal parts of ω and ζ , are to be understood accordingly.

These signal parts are thus nothing else than the quantities representing the *reduction to the reference ellipsoid* in the familiar sense (Heiskanen and Moritz, 1967, secs. 5-4 and 5-5).

It is evident that the expressions (18-11), (18-12), and (18-13) presuppose a geocentric reference ellipsoid; otherwise ξ and η are to be replaced by ξ^a and η^a as given by (18-9), which introduces additional parameters δx_0 , δy_0 , δz_0 .

Thus, in principle, all measurements of α , ω and ζ give information, not only on the geometry through their ellipsoidal parts α' , ω' , and ζ' , but also on the gravity field through their signal parts s_α , s_ω , and s_ζ . Since s_ω is readily seen to be very small, the contribution of ω to the determination of the gravity field will in general be negligible, corresponding to the well-known fact that the reduction of ω to the ellipsoid can usually be neglected (*ibid.*, p.189). On the other hand, s_ζ is significant, which is in agreement with the possibility of using zenith distances for determining deflections of the vertical (*ibid.*, p.176).

Let us finally consider satellite observations. Take, for instance, photographic observations of right ascension α and declination δ , and laser measurements of distances ρ to the satellite. We have relations of the form (cf. *ibid.*, p.355):

$$\alpha = \alpha(x_p, y_p, z_p; t; a_0, e_0, i_0, \Omega_0, \omega_0, T_0; J_{nm}, K_{nm}),$$

$$\delta = \delta(x_P, y_P, z_P; t; a_0, e_0, i_0, \Omega_0, \omega_0, T_0; J_{nm}, K_{nm}) , \quad (18-16)$$

$$\rho = \rho(x_P, y_P, z_P; t; a_0, e_0, i_0, \Omega_0, \omega_0, T_0; J_{nm}, K_{nm}) .$$

The parameter vector X consists of corrections to the station coordinates x_P, y_P, z_P , to the time t , and to the orbital elements $a_0, e_0, i_0, \Omega_0, \omega_0, T_0$; the influence of the reference gravity field is also implicitly contained and may be taken into account by suitable parameters.

The signal parts are represented by the effect of J_{nm} and K_{nm} , the coefficients of the expansion (3-13) of the gravitational potential into spherical harmonics:

$$\begin{aligned} s_\alpha &= \sum_{m,n} \left(\frac{\partial \alpha}{\partial J_{nm}} \delta J_{nm} + \frac{\partial \alpha}{\partial K_{nm}} \delta K_{nm} \right) , \\ s_\delta &= \sum_{m,n} \left(\frac{\partial \delta}{\partial J_{nm}} \delta J_{nm} + \frac{\partial \delta}{\partial K_{nm}} \delta K_{nm} \right) , \\ s_\rho &= \sum_{m,n} \left(\frac{\partial \rho}{\partial J_{nm}} \delta J_{nm} + \frac{\partial \rho}{\partial K_{nm}} \delta K_{nm} \right) , \end{aligned} \quad (18-17)$$

where

$$\delta J_{nm} = J_{nm} - J'_{nm} , \quad \delta K_{nm} = K_{nm} - K'_{nm} = K_{nm} \quad (18-18)$$

are the differences between the actual coefficients J_{nm} and K_{nm} and their normal values J'_{nm} and K'_{nm} referring to the reference gravity field; we have $J'_{nm} = 0$ for $m \neq 0$ and $K'_{nm} = 0$ throughout because of the rotational symmetry of the reference field.

On linearizing (18-16) and taking (18-17) into account we obtain expressions of the form

$$\begin{aligned} \alpha &= \alpha' + s_\alpha , \\ \delta &= \delta' + s_\delta , \\ \rho &= \rho' + s_\rho , \end{aligned} \quad (18-19)$$

again agreeing with our usual model.

Doppler measurements, but also recently proposed observational schemes such as satellite altimetry, satellite-to-satellite tracking, or gradiometry may be treated in precisely the same way. The same holds, e.g. for

the determination of zonal harmonics from variations of the orbital parameters, as we shall see in sec.21.

Systematic errors are taken into account by including them in the vector X ; and to provide for random errors, we add a term n to arrive again at the general model (18-1).

We are thus in a position to apply the basic equations for least-squares collocation. First, the parameters X are obtained from (16-36):

$$\hat{X} = (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} l ; \quad (18-20)$$

then the signal s will be given by (16-37):

$$\hat{s} = C_{st} \bar{C}^{-1} (l - A \hat{X}) . \quad (18-21)$$

Let us recall the meaning of these equations. The vector l comprises all observations of various types as we have just considered; the matrix \bar{C} is the covariance matrix of l . The signal s_p is an arbitrary quantity of the anomalous gravitational field, say a geoidal height at a certain point, a deflection of the vertical at 10 km elevation, or the spherical harmonic coefficient K_{93} . Any field quantity may be obtained in this way by taking the proper covariances C_{st} . As we have seen in sec.11, all signal covariances must be derived from one basic covariance function by appropriate linear operations.

As important limiting cases we have:

Case 1: $X = 0$,

Case 2: $s = 0$,

The limiting case 1, that of no systematic parameters occurring or all such parameters being known, has been the subject of sections 9 to 15; here (18-21) reduces to (14-27):

$$\hat{s} = C_{st} \bar{C}^{-1} l . \quad (18-22)$$

Here l is centered ($\bar{E}\{l\}=0$); if X is not zero but known, we subtract AX from the original observations to get "centered" observations l . For instance, the centered observation corresponding to g is the gravity anomaly $\Delta g = g - \gamma$.

The limiting case 2 corresponds to the absence of the anomalous gravity field. The earth is then considered as an equipotential ellipsoid. Since with $s = 0$ also the signal covariances C_{tt} are zero, the matrix

$$\bar{C} = C_{tt} + C_{nn} = C_{tt} + D \quad (18-23)$$

will consist only of the covariance matrix D of the measuring errors, so that in (18-20) \bar{C} is to be replaced by D :

$$\hat{X} = (A^T D^{-1} A)^{-1} A^T D^{-1} l \quad (18-24)$$

This corresponds to the result of a pure geometrical adjustment in the usual sense, the earth being identified with an ellipsoid.

In a way, our general method splits up the problem into two steps which are very similar to these limiting cases. The first step, the determination of the parameters by (18-20), corresponds to case 2, with the error covariance matrix D replaced by the general covariance matrix \bar{C} , which incorporates also the signal covariances. *The replacement of D by \bar{C} is the only effect of the anomalous gravity field on the determination of the parameters X (which describe, e.g., the geometry).*

Then, using the \hat{X} so obtained, the observations l may be centered by subtracting $A\hat{X}$ to get

$$l' = l - A\hat{X}, \quad (18-25)$$

and then (18-21) gives the signal:

$$\hat{s} = C_{st} \bar{C}^{-1} l' \quad (18-26)$$

This amounts to the use of (18-22), with l replaced by l' .

Properties of the solution. In this way, the signal estimation, that is, the estimation of the anomalous gravitational field, in least-squares collocation with parameters is reduced to the corresponding problem as treated in sections 9 to 15. It has, therefore, basic properties already discussed in these sections, which are related to the invariance and optimality properties pointed out in sections 12 and 16. For instance, we obtain a rigorously consistent gravity field, even if we take a general kernel function $K(P, Q)$ in the place of the covariance function. This field is smooth so as to permit downward continuation and to eliminate spurious irregularities.

If the data are errorless, then they are exactly reproduced (this is the original mathematical idea of collocation!), and different choices of the function $K(P,Q)$ correspond to different possible gravitational fields compatible with the given data. If the data are affected by random measuring errors, then the computed field matches the data not exactly but in such a way that the effect of these errors is kept as small as possible.

Accuracy. The accuracy of the estimated quantities \hat{X} and \hat{s} is expressed by the error covariance matrices (17-44), (17-47), and (17-48). It should be noted that these error covariance matrices are computed only on the basis of the given signal and noise covariances, without needing any real measurements. This fact is common to least-squares adjustment and least-squares collocation.

In adjustment this fact is used, for instance, for the planning of a triangulation: several possible configurations are studied and compared with respect to their accuracy, even before performing any actual measurements. In collocation we can similarly use computed error covariances, for instance, for the planning of gravity surveys (Tscherning, 1975a) and for feasibility studies of methods under development, such as satellite altimetry (Rapp, 1978), aerial gradiometry (Schwarz, 1977), satellite gradiometry (Schwarz and Kryński, 1977), and satellite-to-satellite tracking (Kryński, 1979). Various configurations and types of data (e.g., combinations of satellite and terrestrial data) can be studied in this way.

Accuracy computations presuppose good estimates of the input covariances, in particular a good knowledge of the basic covariance function $K(P,Q)$. This is much more critical here than in the estimation of the quantities X and s themselves. A similar fact is known from least-squares adjustment: the adjusted quantities are rather insensitive with respect to the choice of a-priori weights, whereas a-posteriori accuracies strongly depend on them. However, even with an imperfectly known covariance function, meaningful comparative feasibility studies can be performed. In fact, even if the "absolute" standard errors are not correct, the relation between accuracies obtained may well give useful information of the relative merits of different measurements and configurations.

Computational considerations. It is instructive to compare the evaluation of classical integral formulas such as Stokes' and Vening Meinesz' formulas (sec.2) with least-squares collocation. In the integral formulas, linear operations are performed on the *data*; this must be done numerically and usually involves interpolation procedures. In collocation, linear operations are performed on the *kernel function*, which can be done analytically in a rigorous fashion. Furthermore the operations in collocation are inverse to

those in the integral formulas and usually simpler; for instance, we have differentiations instead of integrations.

In the integral formulas, an inversion is thus, so to speak, built in analytically. In collocation, this must be done numerically and implies inverting the matrix \bar{C} in formulas such as (18-20) and (18-21). This is the main practical problem in collocation since this matrix can be very large.

Collocation is designed to handle simultaneously and combine arbitrary geodetic data of different kind. Integral formulas can handle only data of one type, usually gravity anomalies; for this purpose they remain appropriate and useful. More about integral formulas and collocation will be found in (Neyman, 1974) and (Moritz, 1976a).

As an idealization, we might assume that all geodetic data available at the present time are combined by (18-20) and (18-21) into a single solution for the earth's geometry and gravitational field. As a matter of fact, this cannot be realized in practice because it would involve the inversion of an excessively large matrix \bar{C} .

In practice, the number of data to be combined is limited by the size of matrix that can be inverted by the computer. This implies a suitable representative selection of the data and some working "from the large to the small" in several steps; see the following section.

Computational problems are comparable to those occurring in the adjustment of very large triangulation networks. As in usual adjustment, also in collocation the matrix inversion can frequently be avoided. In the parameterless case we may replace (18-22) by

$$\bar{s} = C_{st} k, \quad (18-27)$$

where the vector k is directly computed by solving the equation

$$\bar{C}k = 1. \quad (18-28)$$

In the presence of parameters we may proceed similarly (Wolf, 1979, p.299).

Finally we point out that the need to invert large matrices (or solve large systems of linear equations) is not restricted to collocation. In fact, *all* possible choices of base functions $\phi_i(P)$ in expressions such as (12-11) lead to $q \times q$ matrices which may not even be symmetrical; cf. (Moritz and Sünkel, 1978, p.28).

Some bibliographical remarks. Least-squares collocation has started from the subject of interpolation of gravity anomalies by least-squares prediction, which is treated, e.g., in Chapter 7 of (Heiskanen and Moritz, 1967),

where also references to the earlier literature are found. The early review articles by Kaula (1963, 1967) are important.

Least-squares prediction of gravity has been adapted from the prediction theory of stochastic processes, a classical reference being (Wiener, 1949). The connection between stochastic processes and kernel functions has been pointed out by Parzen (1961).

The fundamental publication on least-squares collocation as a theory of determining the gravitational field from heterogeneous data is (Krarup, 1969). A comprehensive elementary presentation, which we are partly following in this book, is (Moritz, 1973a). A good picture of the present status is provided by the lecture notes by various authors collected in the volume (Moritz and Sünkel, 1978). Review articles are (Grafarend, 1976), (Moritz, 1978a), (Nash and Jordan, 1978), and (Tscherning, 1978b).

19. STEPWISE COLLOCATION

It is frequently convenient to split up the estimation by collocation into two steps, just as in ordinary least-squares adjustment. This may be done to reduce the size of matrices to be inverted; another application is the use of additional observations to improve the original estimates.

We start with the basic equations of collocation, (16-36) and (16-37):

$$\hat{X} = (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} l, \quad (19-1)$$

$$\hat{s} = C_{st} \bar{C}^{-1} (I - A \hat{X}). \quad (19-2)$$

Again, l is the vector comprising all observations, s is the signal vector to be estimated, X is the vector of parameters to be estimated, \bar{C} is the covariance matrix of the observation vector l , and C_{st} is the covariance matrix between the vectors l and s . The matrix A is the "sensitivity matrix" characterizing the effect of the parameters X on the observed values l according to (16-3),

$$l = AX + t + n. \quad (19-3)$$

Now we split up the observations l into two parts, the first part making up the vector l_1 , and the second part forming the vector l_2 . Thus the observation vector l is partitioned as follows:

$$l = \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} \quad (19-4)$$

(note that l_1 and l_2 are themselves vectors!). The matrices \bar{C} and C_{st} are partitioned accordingly:

$$\bar{C} = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}, \quad (19-5)$$

$$C_{st} = [C_1 \quad C_2]; \quad (19-6)$$

e.g., C_{12} denotes the covariance matrix between the vectors l_1 and l_2 , and C_1 denotes the covariance matrix between the vectors s and l_1 . In the same way we partition the sensitivity matrix:

$$A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \quad (19-7)$$

so that the observation equations (19-3) fall into two parts:

$$l_1 = A_1 X + t_1 + n_1,$$

$$l_2 = A_2 X + t_2 + n_2.$$

Using this partitioning we wish to split up the estimation by (19-1) and (19-2) into two steps. Let us first consider the estimation of the parameters X by (19-1). For this purpose we need the partitioned inverse matrix \bar{C}^{-1} . Writing

$$\bar{C}^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \quad (19-8)$$

we have the well-known relations (cf. Faddeeva, 1959, § 14):

$$B_{22} = (C_{22} - C_{21}C_{11}^{-1}C_{12})^{-1},$$

$$\begin{aligned}
 B_{12} &= -C_{11}^{-1}C_{12}B_{22}, & B_{21} &= -B_{22}C_{21}C_{11}^{-1}, \\
 B_{11} &= C_{11}^{-1} - C_{11}^{-1}C_{12}B_{21} = C_{11}^{-1} + C_{11}^{-1}C_{12}B_{22}C_{21}C_{11}^{-1}.
 \end{aligned}
 \tag{19-9}$$

Using these relations we find

$$\begin{aligned}
 A^T \bar{C}^{-1} A &= \begin{bmatrix} A_1^T & A_2^T \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \\
 &= A_1^T B_{11} A_1 + A_1^T B_{12} A_2 + A_2^T B_{21} A_1 + A_2^T B_{22} A_2 = \\
 &= A_1^T C_{11}^{-1} A_1 + (A_2^T - A_1^T C_{11}^{-1} C_{12}) B_{22} (A_2 - C_{21} C_{11}^{-1} A_1).
 \end{aligned}$$

With the abbreviations

$$\bar{A}_2 = A_2 - C_{21} C_{11}^{-1} A_1, \tag{19-10}$$

$$P_1 = A_1^T C_{11}^{-1} A_1 \tag{19-11}$$

this becomes

$$A^T \bar{C}^{-1} A = P_1 + \bar{A}_2^T B_{22} \bar{A}_2. \tag{19-12}$$

The inverse matrix is then given by

$$(A^T \bar{C}^{-1} A)^{-1} = P_1^{-1} - P_1^{-1} \bar{A}_2^T (B_{22}^{-1} + \bar{A}_2 P_1^{-1} \bar{A}_2^T)^{-1} \bar{A}_2 P_1^{-1}. \tag{19-13}$$

This is readily proved by multiplying the right-hand sides of (19-12) and (19-13): after some straightforward algebra the unit matrix results as it should be.

With the new abbreviation

$$\bar{C}_{22} = C_{22} - C_{21} C_{11}^{-1} C_{12} + \bar{A}_2 P_1^{-1} \bar{A}_2^T \tag{19-14}$$

eq. (19-13) becomes

$$(A^T \bar{C}^{-1} A)^{-1} = P_1^{-1} - P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} \bar{A}_2 P_1^{-1} . \quad (19-15)$$

We further have

$$\begin{aligned} A^T \bar{C}^{-1} l &= \begin{bmatrix} A_1^T & A_2^T \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} l_1 \\ l_2 \end{bmatrix} \\ &= A_1^T B_{11} l_1 + A_1^T B_{12} l_2 + A_2^T B_{21} l_1 + A_2^T B_{22} l_2 , \end{aligned}$$

and substituting (19-9) and using (19-10) we obtain

$$A^T \bar{C}^{-1} l = A_1^T C_{11}^{-1} l_1 + \bar{A}_2^T B_{22} (l_2 - C_{21} C_{11}^{-1} l_1) . \quad (19-16)$$

The substitution of (19-15) and (19-16) into (19-1) gives

$$\begin{aligned} \hat{X} &= (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} l = \\ &= P_1^{-1} A_1^T C_{11}^{-1} l_1 - P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} \bar{A}_2 P_1^{-1} A_1^T C_{11}^{-1} l_1 + \\ &+ (P_1^{-1} - P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} \bar{A}_2 P_1^{-1}) \bar{A}_2^T B_{22} (l_2 - C_{21} C_{11}^{-1} l_1) . \end{aligned} \quad (19-17)$$

The first term on the right-hand side is

$$\hat{X}_1 = P_1^{-1} A_1^T C_{11}^{-1} l_1 = (A_1^T C_{11}^{-1} A_1)^{-1} A_1^T C_{11}^{-1} l_1 , \quad (19-18)$$

which is nothing else but the least-squares collocation estimate of the vector X on the basis of the partial observation vector l_1 only, as the comparison with (19-1) shows.

The last term on the right-hand side may be transformed as follows:

$$\begin{aligned} (P_1^{-1} - P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} \bar{A}_2 P_1^{-1}) \bar{A}_2^T B_{22} &= P_1^{-1} \bar{A}_2^T (I - \bar{C}_{22}^{-1} \bar{A}_2 P_1^{-1} \bar{A}_2^T) B_{22} = \\ &= P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} (\bar{C}_{22} - \bar{A}_2 P_1^{-1} \bar{A}_2^T) B_{22} = P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} B_{22} B_{22} = \\ &= P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} ; \end{aligned} \quad (19-19)$$

here we have used (19-9).

In view of (19-18) and (19-19), eq. (19-17) reduces to

$$\hat{X} = \hat{X}_1 + P_1^{-1} \bar{A}_2^T C_{22}^{-1} (1_2 - C_{21} C_{11}^{-1} 1_1 - \bar{A}_2 \hat{X}_1) . \quad (19-20)$$

This is the required equation for the parameter vector X . The first term on the right-hand side represents the estimate (19-18) on the basis of the first part, 1_1 , of the observations 1 ; the second term expresses the improvement of the estimate by using, in addition, the second part, 1_2 , of the observations.

Now we shall effect a similar transformation for the signal estimate (19-2). Putting

$$\bar{z} = 1 - A\hat{X} , \quad \bar{z} = \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \end{bmatrix} = \begin{bmatrix} 1_1 - A_1 \hat{X} \\ 1_2 - A_2 \hat{X} \end{bmatrix} \quad (19-21)$$

and using (19-5), (19-6), and (19-8) we find

$$\begin{aligned} \bar{s} &= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \begin{bmatrix} \bar{z}_1 \\ \bar{z}_2 \end{bmatrix} = \\ &= (C_1 B_{11} + C_2 B_{21}) \bar{z}_1 + (C_1 B_{12} + C_2 B_{22}) \bar{z}_2 , \end{aligned}$$

whence, by (19-9),

$$\bar{s} = C_1 C_{11}^{-1} \bar{z}_1 + (C_2 - C_1 C_{11}^{-1} C_{12}) B_{22} (\bar{z}_2 - C_{21} C_{11}^{-1} \bar{z}_1) . \quad (19-22)$$

By (19-21) we have

$$\bar{z}_1 = 1_1 - A_1 \hat{X} = 1_1 - A_1 \hat{X}_1 - A_1 (\hat{X} - \hat{X}_1) \quad (19-23)$$

$$\begin{aligned} \bar{z}_2 - C_{21} C_{11}^{-1} \bar{z}_1 &= 1_2 - A_2 \hat{X} - C_{21} C_{11}^{-1} 1_1 + C_{21} C_{11}^{-1} A_1 \hat{X} = \\ &= 1_2 - C_{21} C_{11}^{-1} 1_1 - \bar{A}_2 \hat{X} = \\ &= 1_2 - C_{21} C_{11}^{-1} 1_1 - \bar{A}_2 \hat{X}_1 - \bar{A}_2 (\hat{X} - \hat{X}_1) , \end{aligned} \quad (19-24)$$

using (19-10).
We note that

$$\hat{s}_1 = C_1 C_{11}^{-1} (1_1 - A_1 \hat{x}_1) \quad (19-25)$$

is the estimate of the signal s on the basis of the partial observation vector 1_1 only, as the comparison with (19-2) shows. By means of (19-23), (19-24), and (19-25), eq. (19-22) then becomes

$$\begin{aligned} s &= \hat{s}_1 - C_1 C_{11}^{-1} A_1 (\hat{x} - \hat{x}_1) + \\ &+ (C_2 - C_1 C_{11}^{-1} C_{12}) B_{22} (1_2 - C_{21} C_{11}^{-1} 1_1 - \bar{A}_2 \hat{x}_1) - \\ &- (C_2 - C_1 C_{11}^{-1} C_{12}) B_{22} \bar{A}_2 (\hat{x} - \hat{x}_1) . \end{aligned}$$

The substitution of $\hat{x} - \hat{x}_1$ by (19-20) gives, I denoting the unit matrix,

$$\begin{aligned} \hat{s} &= \hat{s}_1 + \left[(C_2 - C_1 C_{11}^{-1} C_{12}) B_{22} (I - \bar{A}_2 P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1}) - \right. \\ &- \left. C_1 C_{11}^{-1} A_1 P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} \right] (1_2 - C_{21} C_{11}^{-1} 1_1 - \bar{A}_2 \hat{x}_1) \\ &= \hat{s}_1 + R \bar{C}_{22}^{-1} (1_2 - C_{21} C_{11}^{-1} 1_1 - \bar{A}_2 \hat{x}_1) , \end{aligned} \quad (19-26)$$

where we have put

$$R = (C_2 - C_1 C_{11}^{-1} C_{12}) B_{22} (\bar{C}_{22} - \bar{A}_2 P_1^{-1} \bar{A}_2^T) - C_1 C_{11}^{-1} A_1 P_1^{-1} \bar{A}_2^T .$$

By (19-9) and (19-14) we find

$$B_{22} (\bar{C}_{22} - \bar{A}_2 P_1^{-1} \bar{A}_2^T) = I ,$$

so that

$$R = C_2 - C_1 C_{11}^{-1} C_{12} - C_1 C_{11}^{-1} A_1 P_1^{-1} \bar{A}_2^T .$$

Hence (19-26) becomes

$$\mathbf{z} = \mathbf{z}_1 + (\mathbf{C}_2 - \mathbf{C}_1 \mathbf{C}_{11}^{-1} \mathbf{C}_{12} - \mathbf{C}_1 \mathbf{C}_{11}^{-1} \mathbf{A}_1 \mathbf{P}_1^{-1} \bar{\mathbf{A}}_2^T) \bar{\mathbf{C}}_{22}^{-1} (\mathbf{l}_2 - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{l}_1 - \bar{\mathbf{A}}_2 \hat{\mathbf{x}}_1) . \quad (19-27)$$

This is the required equation for stepwise estimation of the signal, together with (19-25). It is analogous to the corresponding equation for $\hat{\mathbf{x}}$ and has the same interpretation.

20. ACCURACY IN STEPWISE COLLOCATION

Using the abbreviations

$$\bar{\mathbf{A}}_2 = \mathbf{A}_2 - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{A}_1 , \quad (20-1)$$

$$\mathbf{P}_1 = \mathbf{A}_1^T \mathbf{C}_{11}^{-1} \mathbf{A}_1 , \quad (20-2)$$

$$\mathbf{T}_2 = \mathbf{l}_2 - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{l}_1 - \bar{\mathbf{A}}_2 \hat{\mathbf{x}}_1 , \quad (20-3)$$

$$\bar{\mathbf{C}}_{22} = \mathbf{C}_{22} - \mathbf{C}_{21} \mathbf{C}_{11}^{-1} \mathbf{C}_{12} + \bar{\mathbf{A}}_2 \mathbf{P}_1^{-1} \bar{\mathbf{A}}_2^T , \quad (20-4)$$

$$\bar{\mathbf{C}}_2 = \mathbf{C}_2 - \mathbf{C}_1 \mathbf{C}_{11}^{-1} \mathbf{C}_{12} - \mathbf{C}_1 \mathbf{C}_{11}^{-1} \mathbf{A}_1 \mathbf{P}_1^{-1} \bar{\mathbf{A}}_2^T , \quad (20-5)$$

and the equations (19-18) and (19-25) for the first step,

$$\hat{\mathbf{x}}_1 = \mathbf{P}_1^{-1} \mathbf{A}_1^T \mathbf{C}_{11}^{-1} \mathbf{l}_1 , \quad (20-6)$$

$$\mathbf{z}_1 = \mathbf{C}_1 \mathbf{C}_{11}^{-1} (\mathbf{l}_1 - \mathbf{A}_1 \hat{\mathbf{x}}_1) , \quad (20-7)$$

we may put the final estimates (19-20) and (19-27) into the form

$$\hat{\mathbf{x}} = \hat{\mathbf{x}}_1 + \mathbf{P}_1^{-1} \bar{\mathbf{A}}_2^T \bar{\mathbf{C}}_{22}^{-1} \mathbf{T}_2 , \quad (20-8)$$

$$\mathbf{z} = \mathbf{z}_1 + \bar{\mathbf{C}}_2 \bar{\mathbf{C}}_{22}^{-1} \mathbf{T}_2 . \quad (20-9)$$

Writing the last equation as

$$\mathbf{z} - \mathbf{z}_1 = \bar{\mathbf{C}}_2 \bar{\mathbf{C}}_{22}^{-1} \mathbf{T}_2 , \quad (20-10)$$

we recognize a striking similarity to the ordinary prediction equation without systematic parameters (9-28):

$$\xi = C_{11} C_{11}^{-1} l_1 . \quad (20-11)$$

Eq. (20-11) holds if the expectation \bar{E} , given by (14-12), of l_1 is zero:

$$\bar{E}\{l_1\} = 0 ;$$

we thus first verify that also

$$\bar{E}\{T_2\} = 0 . \quad (20-12)$$

This follows from (20-3): using (20-1) and (20-6) we have

$$T_2 = l_2 - C_{21} C_{11}^{-1} l_1 - (A_2 - C_{21} C_{11}^{-1} A_1) P_1^{-1} A_1^T C_{11}^{-1} l_1 ; \quad (20-13)$$

and since by (19-3) with (14-14) and (14-15)

$$\bar{E}\{l_1\} = A_1 X , \quad \bar{E}\{l_2\} = A_2 X , \quad (20-14)$$

X denoting the true value of the parameter vector, we find

$$\begin{aligned} \bar{E}\{T_2\} &= \left[A_2 - C_{21} C_{11}^{-1} A_1 - (A_2 - C_{21} C_{11}^{-1} A_1) P_1^{-1} A_1^T C_{11}^{-1} A_1 \right] X \\ &= \left[A_2 - C_{21} C_{11}^{-1} A_1 - (A_2 - C_{21} C_{11}^{-1} A_1) \right] X = 0 , \end{aligned}$$

which proves (20-12).

Introduce now the "true centered observations" z_1 and z_2 by

$$z_1 = l_1 - A_1 X , \quad (20-15)$$

$$z_2 = l_2 - A_2 X ,$$

so that

$$\bar{E}\{z_1\} = 0 , \quad \bar{E}\{z_2\} = 0$$

in view of (20-14). Then (20-13) becomes by means of (20-15),

$$T_2 = z_2 + A_2 X - C_{21} C_{11}^{-1} (z_1 + A_1 X) - (A_2 - C_{21} C_{11}^{-1} A_1) P_1^{-1} A_1^T C_{11}^{-1} (z_1 + A_1 X),$$

which, after obvious cancellations, reduces to

$$T_2 = z_2 - (C_{21} + \bar{A}_2 P_1^{-1} A_1^T) C_{11}^{-1} z_1. \quad (20-16)$$

By means of this expression we readily find

$$\begin{aligned} T_2 T_2^T &= z_2 z_2^T - (C_{21} + \bar{A}_2 P_1^{-1} A_1^T) C_{11}^{-1} z_1 z_2^T - z_2 z_1^T C_{11}^{-1} (C_{12} + A_1 P_1^{-1} \bar{A}_2^T) + \\ &\quad + (C_{21} + \bar{A}_2 P_1^{-1} A_1^T) C_{11}^{-1} z_1 z_1^T C_{11}^{-1} (C_{12} + A_1 P_1^{-1} \bar{A}_2^T). \end{aligned}$$

Forming the average \bar{E} and taking into account that

$$\bar{E}(z_1 z_1^T) = C_{11}, \quad \text{etc.} \quad (20-17)$$

we obtain, after some reduction,

$$\bar{E}\{T_2 T_2^T\} = C_{22} - C_{21} C_{11}^{-1} C_{12} + \bar{A}_2 P_1^{-1} \bar{A}_2^T = \bar{C}_{22},$$

in view of (20-4). This shows that \bar{C}_{22} is indeed the autocovariance matrix of T_2 .

In a similar way it may be verified that \bar{C}_2 is the covariance matrix between $s - \hat{s}_1$ and T_2 : we have

$$(s - \hat{s}_1) \bar{T}_2^T = s \bar{T}_2^T - C_1 C_{11}^{-1} (I_1 - A_1 \hat{X}_1) \bar{T}_2^T.$$

Now,

$$\begin{aligned} I_1 - A_1 \hat{X}_1 &= (I - A_1 P_1^{-1} A_1^T C_{11}^{-1}) I_1 \\ &= (I - A_1 P_1^{-1} A_1^T C_{11}^{-1}) z_1 + (A_1 - A_1 P_1^{-1} A_1^T C_{11}^{-1} A_1) X \\ &= (I - A_1 P_1^{-1} A_1^T C_{11}^{-1}) z_1. \end{aligned}$$

Thus by (20-16)

$$\begin{aligned}(s - \hat{s}_1)^T T_2^T &= s z_2^T - s z_1^T C_{11}^{-1} (C_{12} + A_1 P_1^{-1} \bar{A}_2^T) - \\ &\quad - C_1 C_{11}^{-1} (I - A_1 P_1^{-1} A_1^T C_{11}^{-1}) z_1 z_2^T + \\ &\quad + C_1 C_{11}^{-1} (I - A_1 P_1^{-1} A_1^T C_{11}^{-1}) z_1 z_1^T C_{11}^{-1} (C_{12} + A_1 P_1^{-1} \bar{A}_2^T) .\end{aligned}$$

Forming the average and noting that $E\{s z_1^T\} = C_1$, $E\{s z_2^T\} = C_2$ we readily find that $E\{(s - \hat{s}_1)^T T_2^T\}$ indeed reduces to (20-5), which was to be shown.

Thus (20-10) is, in fact, of the form (20-11). What is more, even (20-8) may be written in this form, namely

$$\bar{X} - \hat{X}_1 = P_1^{-1} \bar{A}_2^T \bar{C}_{22}^{-1} T_2 , \quad (20-18)$$

where $P_1^{-1} \bar{A}_2^T$ represents C_{s1} , in our case the covariance between $X - \hat{X}_1$ and T_2 . The signal s is now $X - \hat{X}_1$ (true value); its estimate is $\hat{X} - \hat{X}_1$. We have

$$X - \hat{X}_1 = X - P_1^{-1} A_1^T C_{11}^{-1} (z_1 + A_1 X) = -P_1^{-1} A_1^T C_{11}^{-1} z_1 ,$$

as the other two terms cancel, which also shows that $E\{X - \hat{X}_1\}$ is zero. We further get

$$(X - \hat{X}_1)^T T_2^T = -P_1^{-1} A_1^T C_{11}^{-1} z_1 z_2^T + P_1^{-1} A_1^T C_{11}^{-1} z_1 z_1^T C_{11}^{-1} (C_{12} + A_1 P_1^{-1} \bar{A}_2^T) .$$

Forming the average \bar{E} we obtain after some reduction

$$\bar{E}\{(X - \hat{X}_1)^T T_2^T\} = P_1^{-1} \bar{A}_2^T ,$$

which was to be shown.

Interpreting both (20-8) and (20-9) as instances of the general prediction formula (20-11) enlarges our theoretical understanding of these formulas, but it is also of considerable practical value since it enables us immediately to write down the corresponding covariances.

The error covariance matrix for the signal vector s estimated by (20-11) is given by (9-29):

$$E_{ss} = C_{ss} - C_{s1}C_{11}^{-1}C_{1s} \quad (20-19)$$

Here C_{ss} represents the "a priori covariance matrix", or "covariance before estimation", and E_{ss} represents the error covariance matrix, or "covariance matrix after estimation". The last term in (20-19) thus represents the gain in accuracy due to the estimation using the observations 1 .

The signal vector s , of which C_{ss} is the covariance matrix, can contain elements of different type since, as we have repeatedly remarked, the present theory is valid for heterogeneous quantities.

Let us now apply (20-19) to our present problem. The signals are now $X - \hat{X}_1$ and $s - \hat{s}_1$, both of which have been found to be random quantities of expectation zero. Their "a priori" covariance matrices

$$\overline{E}\{(s - \hat{s}_1)(s - \hat{s}_1)^T\} = E_{ss,1} \quad (20-20)$$

$$\overline{E}\{(X - \hat{X}_1)(X - \hat{X}_1)^T\} = E_{xx,1} \quad (20-21)$$

$$\overline{E}\{(X - \hat{X}_1)(s - \hat{s}_1)^T\} = E_{xs,1} \quad (20-22)$$

are obviously nothing else than the error covariances after the first step, using the observations 1_1 only. The observations for the second step being T_2 , the covariance matrix C_{11} is now \overline{C}_{22} , as we have already seen, and the matrix C_{s1} is to be taken from (20-10) and (20-18).

On performing these identifications, (20-19) gives

$$E_{ss} = E_{ss,1} - \overline{C}_2 \overline{C}_{22}^{-1} \overline{C}_2^T \quad (20-23)$$

$$E_{xx} = E_{xx,1} - P_1^{-1} \overline{A}_2^T \overline{C}_{22}^{-1} \overline{A}_2 P_1^{-1} \quad (20-24)$$

$$E_{xs} = E_{xs,1} - P_1^{-1} \overline{A}_2^T \overline{C}_{22}^{-1} \overline{C}_2^T \quad (20-25)$$

the last equation results from the fact that an expression analogous to (20-19) also holds for mixed covariances of type (20-22). The error covariances after the first step are given by

$$E_{ss,1} = C_{ss} - C_1 C_{11}^{-1} C_1^T + C_1 C_{11}^{-1} A_1 P_1^{-1} A_1^T C_{11}^{-1} C_1^T \quad (20-26)$$

$$E_{xx,1} = P_1^{-1} \quad (20-27)$$

$$E_{x\bar{x},1} = -P_1^{-1}A_1^T C_{11}^{-1}C_1^T. \quad (20-28)$$

This follows directly from (17-47), (17-44), and (17-48). The notations (20-1) to (20-5) have been used.

As a matter of fact, the expressions (20-23), (20-24) and (20-25) could also be obtained by partitioning the original equations, in the same way as expressions for \bar{X} and \bar{S} were obtained in the last section. This is straightforward for (20-24)--this expression is a direct consequence of (19-15)--; it is not too difficult for (20-25); but it is rather laborious for (20-23).

Equations (20-8), (20-9), (20-23), (20-24), and (20-25) are the basic formulas for performing the collocation in two steps. Two applications immediately present themselves:

1. If the size of the matrix \bar{C} to be inverted is too large, then a stepwise procedure might be feasible since C_{11} and \bar{C}_{22} are smaller matrices than \bar{C} .
2. Let signals and parameters have been estimated using observations l_1 . If new observations l_2 are available, the original estimates can be improved in this way.

It is also possible to increase the number of observations by only one at a time, repeating this procedure in a successive manner; thus l_2 consists of one observation only. This may be called *sequential collocation*.

As a simple example, let there be 100 measurements l_i ($i = 1, 2, \dots, 100$; here l_i are numbers, not vectors), and let the vector X consist of 10 parameters. As a first step, compute \bar{X} and the desired signals from an original data set of, say, the first 15 observations (the minimum would be 10 observations). As a second step, use the sixteenth observation considered as a one-component vector l_2 , using the basic formulas for stepwise collocation. As the next step, consider the seventeenth observations as l_2 and so on, up to the hundredth.

It is remarkable in this method that, except for the computation of a first estimate using the original data set, no matrix inversion is needed any more. In fact, \bar{C}_{22} now reduces to a single element, so that its inverse is simply the reciprocal of that element, and P_1^{-1} is computed successively from (20-24) since E_{xx} is P_1^{-1} for the next step.

This advantage is, however, made up by a great increase of matrix multiplications, so that the total computational effort is, in general, not reduced. In fact, stepwise collocation may be compared to matrix inversion by partitioning, and sequential collocation then corresponds to matrix inversion by bordering (Faddeeva, 1959, §14 and 15).

Sequential adjustment (Deutsch, 1965, chapter 8) may obviously be considered as a special case of sequential collocation, and Kalman filtering is known to be closely related to sequential adjustment (*ibid.*, chapter 12). The main differences between Kalman filtering and sequential collocation are:

1. In sequential collocation, the system parameters X do not change, whereas in Kalman filtering they undergo a linear transformation perturbed by internal noise.

2. In sequential collocation, but not in Kalman filtering, we have signal parameters related to the observations only through their covariances. This is characteristic of collocation; see the corresponding remarks in sec.16.

A general operational program for stepwise least-squares collocation has been given by Tscherning (1974).

21. DETERMINATION OF SPHERICAL HARMONICS

Least-squares collocation may also be applied to the determination of spherical harmonics of the anomalous gravitational field from satellite observations. This application is of particular practical importance and provides an opportunity to look at collocation from a different angle.

Consider, for example, an observed distance ρ to a satellite. Then the third equation of (18-17) may be written in the form

$$s_{\rho} = \sum_{r=1}^{\infty} B_r s_r \quad (21-1)$$

Here s_1, s_2, \dots represent the spherical-harmonic coefficients $\delta \bar{J}_{nm}$ and $\delta \bar{K}_{nm}$ (it is convenient to use fully normalized harmonics), arranged in some way as a linear sequence. For instance, we may take

$$\begin{aligned} s_1 &= \delta \bar{J}_{20}, & s_2 &= \delta \bar{J}_{21}, & s_3 &= \delta \bar{K}_{21}, & s_4 &= \delta \bar{J}_{22}, \\ s_5 &= \delta \bar{K}_{22}, & s_6 &= \delta \bar{J}_{30}, & s_7 &= \delta \bar{J}_{31}, & s_8 &= \delta \bar{K}_{31}, \\ s_9 &= \delta \bar{J}_{32}, & s_{10} &= \delta \bar{K}_{32}, & s_{11} &= \delta \bar{J}_{33}, & s_{12} &= \delta \bar{K}_{33}, \\ s_{13} &= \delta \bar{J}_{40}, & s_{14} &= \delta \bar{J}_{41}, & & & & \dots \end{aligned} \quad (21-2)$$

The coefficients B_r are the corresponding quantities $\partial \rho / \partial J_{nm}$ or $\partial \rho / \partial K_{nm}$, as the case may be. For instance, according to the ordering just given, we have

$$B_3 = \frac{\partial \rho}{\partial K_{21}} \quad \text{and} \quad B_4 = \frac{\partial \rho}{\partial J_{22}} .$$

Each observed distance, but also any other satellite observation, gives an equation of type (21-1), which represents the expansion of the measured quantity into a series of spherical harmonics. Assuming q observations, we thus get

$$\begin{aligned} t_1 &= B_{11}s_1 + B_{12}s_2 + B_{13}s_3 + \dots , \\ t_2 &= B_{21}s_1 + B_{22}s_2 + B_{23}s_3 + \dots , \\ &\vdots \\ t_q &= B_{q1}s_1 + B_{q2}s_2 + B_{q3}s_3 + \dots , \end{aligned} \tag{21-3}$$

where t_i represents the "signal part" of the i -th observation.

In matrix notation we may write the linear system (21-3) in the form

$$t = Bs , \tag{21-4}$$

where t is the (column) vector

$$t = [t_1 \quad t_2 \quad \dots \quad t_q]^T \tag{21-5}$$

and s is the vector

$$s = [s_1 \quad s_2 \quad s_3 \quad \dots]^T . \tag{21-6}$$

Note that s is an infinite vector, comprising infinitely many components, corresponding to the fact that there are infinitely many spherical harmonics. The symbol B denotes a " $q \times \infty$ matrix":

$$B = \begin{bmatrix} L_{11} & L_{12} & L_{13} & \cdots \\ L_{21} & L_{22} & L_{23} & \cdots \\ \vdots & \vdots & \vdots & \\ L_{q1} & L_{q2} & L_{q3} & \cdots \end{bmatrix} . \quad (21-7)$$

If the observations include also systematic parameters, then we may again use the observation equations (18-1)

$$l = AX + t + n \quad (21-8)$$

and the solution (18-20) and (18-21):

$$\hat{X} = (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} l , \quad (21-9)$$

$$\hat{s} = C_{st} \bar{C}^{-1} (l - A\hat{X}) . \quad (21-10)$$

In this way we can estimate arbitrary elements of the anomalous gravitational field, that is, an arbitrary signal vector s . The present problem is the determination of the spherical-harmonic coefficients of the anomalous gravitational potential; hence our signal vector s is the vector (21-6). It is true that in practice we do not determine the whole infinite vector (21-6) but only the terms up to a certain degree and order. However, since any signal is obtained independently of the others (p.167), there is no harm in first seeking the whole infinite vector s (in the general formula this is even simpler); later on we may numerically evaluate only those terms that interest us.

Thus our vectors s and t are given by (21-6) and (21-5). Let us now compute the covariances that are required in (21-9) and (21-10). By (16-29) we have

$$\bar{C} = C_{tt} + C_{nn} = C_{tt} + D . \quad (21-11)$$

The noise covariance matrix D characterizes the statistical behavior of the measuring errors of the observations. It is a $q \times q$ matrix and is assumed to be given as usual. The signal covariance matrices must be computed by covariance propagation as shown below.

The basic covariance matrix. Let us represent the anomalous potential T on the sphere $r = R$ as a series of fully normalized harmonics (10-6):

$$T(\theta, \lambda) = \sum_{n=2}^{\infty} \sum_{m=0}^n \left[\bar{a}_{nm} \bar{R}_{nm}(\theta, \lambda) + \bar{b}_{nm} \bar{S}_{nm}(\theta, \lambda) \right], \quad (21-12)$$

In the region $r \geq R$ we then have

$$T(r, \theta, \lambda) = \sum_{n=2}^{\infty} \sum_{m=0}^n \left(\frac{R}{r} \right)^{n+1} \left[\bar{a}_{nm} \bar{R}_{nm}(\theta, \lambda) + \bar{b}_{nm} \bar{S}_{nm}(\theta, \lambda) \right], \quad (21-13)$$

since this expression represents the (unique) function harmonic outside the sphere $r = R$ and reducing to (21-12) at the surface of this sphere.

The spherical-harmonic coefficients are given by (3-29):

$$\begin{aligned} \bar{a}_{nm} &= \frac{1}{4\pi} \iint_{\sigma} T(\theta, \lambda) \bar{R}_{nm}(\theta, \lambda) d\sigma, \\ \bar{b}_{nm} &= \frac{1}{4\pi} \iint_{\sigma} T(\theta, \lambda) \bar{S}_{nm}(\theta, \lambda) d\sigma, \end{aligned} \quad (21-14)$$

these expressions are linear functionals of the potential T .

The covariances between these coefficients are thus found by covariance propagation; cf. p. 87. The application of (11-14) yields

$$\begin{aligned} \text{cov}(\bar{a}_{nm}, \bar{a}_{qp}) &= \frac{1}{16\pi^2} \iint_{\sigma} \iint_{\sigma'} K(P, Q) \bar{R}_{nm}(\theta, \lambda) \cdot \\ &\quad \cdot \bar{R}_{qp}(\theta', \lambda') d\sigma d\sigma', \end{aligned} \quad (21-15)$$

where P and Q are points on the sphere $r = R$ with coordinates (θ, λ) and (θ', λ') , respectively. The integrals are defined by

$$\begin{aligned} \iint_{\sigma} d\sigma &= \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} \sin\theta d\theta d\lambda, \\ \iint_{\sigma'} d\sigma' &= \int_{\lambda'=0}^{2\pi} \int_{\theta'=0}^{\pi} \sin\theta' d\theta' d\lambda'. \end{aligned} \quad (21-16)$$

The function $K(P, Q)$ is given by (10-7), in which we express $P_n(\cos\psi)$ by (3-30). We further replace the indices n and m by s and r . Thus we obtain

$$K(P, Q) = K(\psi) = \sum_{s=2}^{\infty} k_s P_s(\cos \psi) \\ = \sum_{s=2}^{\infty} \sum_{r=0}^s \frac{k_s}{2s+1} \left[\bar{R}_{sr}(\theta, \lambda) \bar{R}_{sr}(\theta', \lambda') + \bar{S}_{sr}(\theta, \lambda) \bar{S}_{sr}(\theta', \lambda') \right].$$

We substitute this into (21-15) and interchange summation and integration, getting

$$\text{cov}(\bar{a}_{nm}, \bar{a}_{qp}) = \sum_{s=2}^{\infty} \sum_{r=0}^s \frac{k_s}{2s+1} \cdot \\ \cdot \left[\frac{1}{4\pi} \iint_{\sigma} \bar{R}_{nm}(\theta, \lambda) \bar{R}_{sr}(\theta, \lambda) d\sigma \cdot \frac{1}{4\pi} \iint_{\sigma'} \bar{R}_{qp}(\theta', \lambda') \bar{R}_{sr}(\theta', \lambda') d\sigma' + \right. \\ \left. + \frac{1}{4\pi} \iint_{\sigma} \bar{R}_{nm}(\theta, \lambda) \bar{S}_{sr}(\theta, \lambda) d\sigma \cdot \frac{1}{4\pi} \iint_{\sigma'} \bar{R}_{qp}(\theta', \lambda') \bar{S}_{sr}(\theta', \lambda') d\sigma' \right].$$

Because of the orthogonality relations (3-16) all integrals are zero except the first one if $s = n$, $r = m$ and the second if $s = q$, $r = p$, which are then given by (3-27). Thus we get

$$\text{cov}(\bar{a}_{nm}, \bar{a}_{nm}) = \frac{k_n}{2n+1}, \quad (21-17)$$

$$\text{cov}(\bar{a}_{nm}, \bar{a}_{qp}) = 0 \quad \text{if} \quad n \neq q \quad \text{or} \quad m \neq p \quad \text{or both.}$$

Similarly we find

$$\text{cov}(\bar{b}_{nm}, \bar{b}_{nm}) = \frac{k_n}{2n+1},$$

$$\text{cov}(\bar{b}_{nm}, \bar{b}_{qp}) = 0 \quad \text{if} \quad n \neq q \quad \text{or} \quad m \neq p \quad \text{or both}, \quad (21-18)$$

$$\text{cov}(\bar{a}_{nm}, \bar{b}_{qp}) = 0 \quad \text{always}.$$

If the gravitational potential V is represented in the form (3-13) but with fully normalized instead of conventional harmonics, then

$$T(r, \theta, \lambda) = V - V_{\text{Ellipsoid}} \\ = - \frac{GM}{r} \sum_{n=2}^{\infty} \sum_{m=0}^n \left(\frac{a}{r} \right)^n \left[\delta J_{nm} \bar{R}_{nm}(\theta, \lambda) + \delta K_{nm} \bar{S}_{nm}(\theta, \lambda) \right]. \quad (21-19)$$

The comparison with (21-12) shows that, as a spherical approximation ($r \approx R \approx a$),

$$\delta J_{nm} = -\frac{R}{GM} \bar{a}_{nm}, \quad \delta \bar{K}_{nm} = -\frac{R}{GM} \bar{b}_{nm}. \quad (21-20)$$

from (21-17) and (21-18) we thus finally get

$$\text{cov}(\delta J_{nm}, \delta J_{nm}) = \text{cov}(\delta \bar{K}_{nm}, \delta \bar{K}_{nm}) = \left(\frac{R}{GM}\right)^2 \frac{k_n}{2n+1}, \quad (21-21)$$

all other covariances between spherical-harmonic coefficients being zero. Since the coefficients δJ_{nm} and $\delta \bar{K}_{nm}$ together form the infinite vector s by (21-2), the signal covariance matrix K of s , defined by

$$K = \text{cov}(s, s) = \overline{E}\{ss^T\} = M\{ss^T\}, \quad (21-22)$$

is an infinite diagonal matrix:

$$K = \begin{bmatrix} k_{11} & & & & \\ & k_{22} & & & \\ & & k_{33} & & \\ & & & k_{44} & \\ & 0 & & & \ddots \end{bmatrix}. \quad (21-23)$$

Furthermore, the k_{ii} for all $2n+1$ coefficients of the same degree n are equal, as (21-21) shows. For the ordering (21-2) we get

$$\begin{aligned} k_{11} &= k_{22} = \dots = k_{55} = \left(\frac{R}{GM}\right)^2 \cdot \frac{1}{5} k_2, \\ k_{66} &= k_{77} = \dots = k_{12,12} = \left(\frac{R}{GM}\right)^2 \cdot \frac{1}{7} k_3, \\ k_{13,13} &= k_{14,14} = \dots = k_{21,21} = \left(\frac{R}{GM}\right)^2 \cdot \frac{1}{9} k_4 \end{aligned} \quad (21-24)$$

* * * * *

The reader should distinguish the k_{11} , the diagonal elements in the matrix (21-23), from the k_n , the coefficients in the series (10-7) for the basic covariance function.

The covariance matrix K can thus be easily expressed in terms of the coefficients k_n of the covariance function $K(P,Q)$; it will be called the basic covariance matrix.

Derived signal covariance matrices. From the basic matrix

$$K = C_{ss} \quad (21-25)$$

we may easily derive the other signal covariance matrices C_{st} and C_{tt} , entering into (21-9), (21-10), and (21-11). In fact, t is related to s by (21-4):

$$t = Bs ,$$

which has the form (11-16), so that (11-17) and (11-18) immediately give

$$C_{st} = \text{cov}(s,t) = KB^T , \quad (21-26)$$

$$C_{tt} = \text{cov}(t,t) = BKB^T . \quad (21-27)$$

It should, however, be noted that here K and B are infinite matrices: K is an $\infty \times \infty$ matrix and B is a $q \times \infty$ matrix. With such infinite matrices we can operate exactly as with finite matrices, provided the sums, which are now infinite series, converge. (The mathematically-minded reader may try himself to find a rigorous justification of this and other formal manipulations!).

It is clear that C_{st} is a $\infty \times q$ matrix and C_{tt} is a $q \times q$ matrix. By (21-11) and (21-27) we have

$$\bar{C} = BKB^T + D . \quad (21-28)$$

Having thus found all covariances, we are now ready to evaluate the estimation equations (21-9) and (21-10) and also the corresponding error covariance matrices for the solution, as given in sec.17.

The parameterless case. Of particular importance is the special case in which the parameters are considered known or have been determined before. Then the effect of the parameters is subtracted from the original observation vector and the result is taken as our new observation vector l , so

that (21-8) reduces to

$$l = t + n = Bs + n . \quad (21-29)$$

The solution (21-10) becomes with (21-26) and (21-28) and with $A = 0$:

$$\hat{s} = KB^T(BKB^T + D)^{-1}l . \quad (21-30)$$

There is an interesting relation to the theory of generalized matrix inverses. If we assume errorless observations ($n = 0$), then the system of observation equations (21-29) reduces to

$$Bs = l , \quad (21-31)$$

or explicitly

$$\begin{aligned} B_{11}s_1 + B_{12}s_2 + B_{13}s_3 + \dots &= l_1 , \\ B_{21}s_1 + B_{22}s_2 + B_{23}s_3 + \dots &= l_2 , \\ &\vdots \\ B_{q1}s_1 + B_{q2}s_2 + B_{q3}s_3 + \dots &= l_q , \end{aligned} \quad (21-32)$$

and the problem is to solve this system for the vector s . Corresponding to (21-30), with $n = 0$, we have the solution (for \hat{s} we write simply s)

$$s = KB^T(BKB^T)^{-1}l , \quad (21-33)$$

which, in fact, satisfies (21-31): we have

$$Bs = BKB^T(BKB^T)^{-1}l = l ; \quad (21-34)$$

BKB^T is a $q \times q$ matrix assumed to be regular.

It is clear that (21-33) gives a solution of the system (21-31) for arbitrary matrices K , especially for arbitrary positive-definite matrices K , provided the occurring infinite sums converge and the matrix BKB^T is regular. In fact, the system (21-31) is an underdetermined system of linear equations, which has infinitely many solution vectors s ; each choice of

K gives a possible solution vector, all of them reproducing exactly the measurements l in view of (21-34).

If the solution of (21-31) is formally written as

$$s = B^{-1} l, \quad (21-35)$$

then B^{-1} is a *generalized inverse* (in the sense of A. Bjerhammar) of the rectangular matrix B , and the comparison with (21-33) shows that

$$B^{-1} = KB^T(BKB^T)^{-1}; \quad (21-36)$$

according to (Bjerhammar, 1973, p.106) the inverse B^{-1} may indeed be represented in this way; the fact that B is now an infinite matrix makes formally no difference.

The solution (21-33) satisfies the minimum condition

$$sK^{-1}s^T = \text{minimum}. \quad (21-37)$$

For finite matrices K this is well-known from the theory of generalized inverses (*ibid.*, p.116); for the present case of an infinite matrix K we must presuppose convergence as usual.

On the other hand, the solution (21-30) satisfies the condition

$$sK^{-1}s^T + nD^{-1}n^T = \text{minimum}, \quad (21-38)$$

according to (16-15). It does not satisfy (21-31) but (21-29), so that random measuring errors are taken into account.

Accuracy. The accuracy of the signal vector s is given by (17-45). For the parameterless case we put $A = 0$, obtaining

$$E_{ss} = C_{ss} - C_{st} \bar{C}^{-1} C_{ts}. \quad (21-39)$$

Here

$$C_{ss} = K \quad (21-40)$$

by (21-22); the other covariance matrices are given by (21-26) and (21-28), and $C_{ts} = C_{st}^T$.

An elegant interpretation of E_{ss} has been given by Burkhard and Jackson (1976, p.1514). Let us write (21-30) in the form

$$\hat{s} = L l \quad (21-41)$$

where

$$L = K B^T (B K B^T + D)^{-1} \quad (21-42)$$

We substitute this into (21-29),

$$l = B s + n ,$$

in which s and n denote the "true" values. The result is

$$\hat{s} = L B s + L n , \quad (21-43)$$

so that the "true error" of the estimate value \hat{s} is given by

$$\hat{s} - s = (L B - I) s + L n , \quad (21-44)$$

I denoting the unit matrix as usual. The first term on the right hand side,

$$e_1 = (L B - I) s , \quad (21-45)$$

denotes the *resolving error*, due to the deviation of $L B$ from the unit matrix: if the system (21-31) could be uniquely solved--if B were a regular square matrix--then e_1 would be zero. The second term,

$$e_2 = L n , \quad (21-46)$$

expresses simply the effect of data errors propagating into the solution.

The covariance matrices of e_1 and e_2 are

$$E_1 = \overline{E}\{e_1 e_1^T\} = (L B - I) K (L B - I)^T , \quad (21-47)$$

$$E_2 = \overline{E}\{e_2 e_2^T\} = L D L^T . \quad (21-48)$$

Now

$$E_{ss} = \overline{E}\{(\hat{s} - s)(\hat{s} - s)^T\} \quad (21-49)$$

becomes simply

$$E_{ss} = E_1 + E_2 \quad (21-50)$$

in view of (21-44), e_1 and e_2 being uncorrelated since s and n are uncorrelated. It is straightforward to verify that (21-50) is identical to (21-39).

Overdetermined systems. Let now B , instead of a $q \times \infty$ matrix, be a $q \times N$ matrix, with $N < q$. It is thus a finite "standing" matrix (more rows than columns), instead of a "lying" infinite matrix. The system (21-31) is then overdetermined and has, in general, no solution. However, the system (21-29) can still be solved since the vector n takes care of the necessary flexibility. Now s is an N -vector and K is an $N \times N$ matrix.

The condition (21-38) gives a solution that is formally identical to (21-30). The new feature, peculiar to the overdetermined case $q > N$, is now that the following well-known matrix identity can be applied:

$$KB^T(BKB^T + D)^{-1} = (B^TD^{-1}B + K^{-1})^{-1}B^TD^{-1} . \quad (21-51)$$

This is immediately verified:

$$(B^TD^{-1}B + K^{-1})KB^T = B^TD^{-1}(BKB^T + D) ,$$

$$B^TD^{-1}BKB^T + B^T = B^TD^{-1}BKB^T + B^T .$$

Thus, for the overdetermined case, the solution (21-30) is equivalent to

$$\hat{s} = (B^TD^{-1}B + K^{-1})^{-1}B^TD^{-1}l . \quad (21-52)$$

Note that this formula requires the inversion of an $N \times N$ matrix, whereas (21-30) involves the inversion of a $q \times q$ matrix.

A similar reduction for the full model

$$l = AX + Bs + n ,$$

when B is again a $q \times N$ matrix, has been given by Schwarz (1976a, 1978a).

In the special case that K is a multiple of the $N \times N$ unit matrix I ,

$$K = \lambda I , \quad (21-53)$$

eq. (21-52) becomes

$$\hat{s} = (B^T D^{-1} B + \lambda^{-1} I)^{-1} B^T D^{-1} l, \quad (21-54)$$

and if we let

$$\lambda \rightarrow \infty, \quad (21-55)$$

we get

$$\hat{s} = (B^T D^{-1} B)^{-1} B^T D^{-1} l, \quad (21-56)$$

which is the well-known solution of the overdetermined system

$$l = B s + n, \quad (21-57)$$

by least-squares adjustment by parameters under the condition

$$n^T D^{-1} n = \text{minimum}, \quad (21-58)$$

the s being considered as free functional parameters.

Practical application. The mathematical model discussed in the present section, especially in the parameterless form, has been applied frequently for the determination of spherical harmonics (zonal and others) from satellite observations; cf. (Moritz and Schwarz, 1973), (Schwarz, 1975, 1978a), (Balmino et al., 1976), (Lerch et al., 1977), or (Kostelecký and Klokočník, 1978).

A property of the present solution is that any component of the infinite vector s is obtained by (21-30) independently of the other: the vector $(B K B^T + D)^{-1} l$ is the same for all elements, the r -th element s_r being found by multiplying the r -th row of $K B^T$ by this vector. Thus we may restrict our computation, say, to the first 50 elements of s_r or to the first 20 zonal harmonics only.

It often happens that the series occurring in (21-32) converge quite rapidly, so that the terms $B_{ir} s_r$ can be neglected for $r > N$. If the number of measurements $q > N$, then we have the overdetermined case considered above, and (21-30) can be replaced by (21-52), which involves the inversion of a smaller matrix.

These solutions, corresponding to the minimum principle (21-38), have the usual properties of least-squares collocation, such as minimum variance, but also favorable numerical properties, in particular stability, in comparison to solutions such as (21-33) satisfying (21-37) or (21-56) satisfying (21-58). These latter solutions may be regarded as extreme cases of the general solutions (21-30) or (21-52). For a proper choice of K and D , the general solutions satisfying (21-38) seem to strike a good balance between the extremes.

It also turns out that (21-39) provides reasonable accuracy estimates and also a means to decide which spherical-harmonic coefficients can be meaningfully determined from the data. In fact, C_{ss} represents the "a priori" covariance matrix of s , corresponding to the case that we take $\hat{s} = 0$ in lack of better information. On the other hand, E_{ss} is the "a posteriori" covariance matrix, corresponding to the case that we take \hat{s} to be the least-squares collocation estimate. Then the term $-C_{st} \bar{C}^{-1} C_{ts}$ expresses the gain in accuracy due to the observations used.

In order to compare E_{ss} with C_{ss} , we transform both matrices simultaneously to a diagonal form, in such a way that C_{ss} becomes the unit matrix I :

$$\begin{aligned} UC_{ss}U^T &= I, \\ UE_{ss}U^T &= \Lambda. \end{aligned} \tag{21-59}$$

This can always be achieved by a suitable choice of the matrix U (in mathematical terms, we have a general eigenvalue problem). The matrix Λ is a diagonal matrix, and we denote the diagonal elements by $\lambda_1, \lambda_2, \lambda_3, \dots$. It may be shown that $\lambda_r \leq 1$. If

$$\lambda_r \ll 1, \tag{21-60}$$

then the r -th element u_r of the transformed signal

$$u = Us, \tag{21-61}$$

which is a linear combination of spherical harmonics, is well determined. For more details see (Schwarz, 1978a).

A similar model (of overdetermined type) has been applied to the combination of satellite-determined harmonic coefficients with surface gravity data (Rapp, 1973, 1975).

Finally we mention geophysical inverse problems, which deal with the determination of the inner structure of the earth from measurements performed at the earth's surface. These measurements can never determine the inner structure in a unique manner, similarly as the geodetic measurements cannot fully and uniquely determine the gravitational field. Geophysical inverse problems can again be reduced to systems of linear equations of the form (21-31), and solutions of form (21-30) have been suggested for them. Cf. (Burkhard and Jackson, 1976); a general Hilbert space approach has been given by Backus (1970).

22. LOCAL STRUCTURE OF COVARIANCE FUNCTIONS

The covariance function $K(P,Q)$ of the anomalous potential T has been chosen as the basic covariance function, or kernel function, from which all other signal covariances are derived by covariance propagation. This choice of $K(P,Q)$ is justified by the fact that there are simple relations between T and derived signal quantities, such as Δg , ξ , n , etc., which entail simple relations between $K(P,Q)$ and derived signal covariances, as we have seen, for instance, in sec.15.

On the other hand, from the point of view of empirical determination, the covariance function $C(P,Q)$ of the gravity anomaly Δg has a more fundamental character because gravity anomalies form the main empirical material for the practical determination of the signal covariances.

There is, however, no difficulty in working either with $K(P,Q)$ or with $C(P,Q)$, according to the purpose we have in mind. In fact, there is a one-to-one correspondence between the two functions, as we have seen in sec.13. Let $K(P,Q)$ and $C(P,Q)$ have the spherical-harmonic expansions

$$K(P,Q) = \sum_{n=2}^{\infty} k_n \left(\frac{R^2}{rr'} \right)^{n+1} P_n(\cos\psi) , \quad (22-1)$$

$$C(P,Q) = \sum_{n=2}^{\infty} c_n \left(\frac{R^2}{rr'} \right)^{n+2} P_n(\cos\psi) , \quad (22-2)$$

then the coefficients are related simply by

$$c_n = \left(\frac{n-1}{R} \right)^2 k_n . \quad (22-3)$$

Planar approximation. In this section we shall examine more closely both the local and the global structure of covariance functions. First we shall have a look at the local structure of the function $K(P,Q)$. In a neighborhood of a certain point of the sphere, the spherical surface may approximately be replaced by its tangent plane, which is taken as the xy -plane ($z = 0$) of a local cartesian coordinate system. Then, in this plane, a homogeneous and isotropic covariance function will be a function only of the distance

$$\rho = \sqrt{(x'-x)^2 + (y'-y)^2} \quad (22-4)$$

between the points $P = (x,y,0)$ and $Q = (x',y',0)$:

$$K(P,Q) = K(\rho) . \quad (22-5)$$

Homogeneity means invariance with respect to translation, and isotropy means invariance with respect to rotation; the first property is expressed by the fact that C depends only on coordinate differences $x'-x$ and $y'-y$, and the second property holds because the azimuth does not enter in (22-5).

If we now extend the function (22-5) into outer space, that is to values $z > 0$, then the covariance function will have the form

$$K(P,Q) = K(\rho, z, z') , \quad (22-6)$$

if $P = (x,y,z)$ and $Q = (x',y',z')$, ρ being again the horizontal distance given by (22-4).

It is not difficult to see that the dependence on z and z' can only be through the sum $z+z'$. In fact, K must be harmonic as a function both of P and of Q :

$$\begin{aligned} \frac{\partial^2 K}{\partial x^2} + \frac{\partial^2 K}{\partial y^2} + \frac{\partial^2 K}{\partial z^2} &= 0 , \\ \frac{\partial^2 K}{\partial x'^2} + \frac{\partial^2 K}{\partial y'^2} + \frac{\partial^2 K}{\partial z'^2} &= 0 . \end{aligned} \quad (22-7)$$

In view of (22-4) it is straightforward to verify that

$$\frac{\partial^2 K}{\partial x'^2} = \frac{\partial^2 K}{\partial x^2} , \quad \frac{\partial^2 K}{\partial y'^2} = \frac{\partial^2 K}{\partial y^2} ; \quad (22-8)$$

this is done in the same way in which eq. (22-16) below is derived. Therefore it follows from (22-7) that also

$$\frac{\partial^2 K}{\partial z'^2} = \frac{\partial^2 K}{\partial z^2} . \quad (22-9)$$

The general solution of this partial differential equation is well-known to be (cf. Courant and Hilbert, 1962, p.6)

$$K(P,Q) = F(\rho, z+z') + f(\rho, z-z') , \quad (22-10)$$

with arbitrary (twice differentiable) functions F and f . The planar approximation of (22-1) always leads to functions that depend on $z+z'$ since, putting $r = R + z$, $r' = R + z'$,

$$\frac{R^2}{rr'} = \left(1 + \frac{z}{R}\right)^{-1} \left(1 + \frac{z'}{R}\right)^{-1} \doteq 1 - \frac{z+z'}{R} . \quad (22-11)$$

This rules out functions of type f , so that K must have the form

$$K(P,Q) = F(\rho, Z) , \quad (22-12)$$

where

$$Z = z + z' , \quad (22-13)$$

which was to be shown.

Gradient covariances. From the form (22-12) we may derive important consequences for the covariances of first order gradients T_x , T_y , T_z , where, e.g.,

$$T_x = \frac{\partial T}{\partial x} .$$

The covariances of these gradients are readily expressed in terms of $K(P,Q)$ by covariance propagation (sec.11). We have

$$\text{cov}(T_x, T'_x) = \frac{\partial^2 K}{\partial x \partial x'} ,$$

$$\text{cov}(T_y, T'_y) = \frac{\partial^2 K}{\partial y \partial y^T} \quad (22-14)$$

$$\text{cov}(T_z, T'_z) = \frac{\partial^2 K}{\partial z \partial z^T} \quad (22-15)$$

where

$$T_x = \left(\frac{\partial T}{\partial x} \right)_P, \quad T'_x = \left(\frac{\partial T}{\partial x} \right)_Q \quad (22-15)$$

and similarly for the other gradients.

The differentiation of (22-12) gives

$$\frac{\partial K}{\partial x} = \frac{\partial F}{\partial \rho} \frac{\partial \rho}{\partial x} = - \frac{\partial F}{\partial \rho} \frac{x' - x}{\rho}$$

by (22-4), and further

$$\frac{\partial^2 K}{\partial x^2} = \frac{1}{\rho} \frac{\partial F}{\partial \rho} + \frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\frac{1}{\rho} \frac{\partial F}{\partial \rho} \right) (x' - x)^2 = - \frac{\partial^2 K}{\partial x \partial x^T} \quad (22-16)$$

In the same way we find

$$\frac{\partial^2 K}{\partial y^2} = \frac{1}{\rho} \frac{\partial F}{\partial \rho} + \frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\frac{1}{\rho} \frac{\partial F}{\partial \rho} \right) (y' - y)^2 = - \frac{\partial^2 K}{\partial y \partial y^T} \quad (22-17)$$

We further have by (22-13)

$$\frac{\partial^2 K}{\partial z^2} = \frac{\partial^2 K}{\partial z^2} = \frac{\partial^2 K}{\partial z \partial z^T} \quad (22-18)$$

Now Laplace's equation (22-7), first equation, gives immediately

$$\frac{\partial^2 K}{\partial z \partial z^T} = \frac{\partial^2 K}{\partial x \partial x^T} + \frac{\partial^2 K}{\partial y \partial y^T} \quad (22-19)$$

which provides an important relation between the covariance functions of the first-order gradients T_x, T_y, T_z . This relation is rather surprising since a similar relation between the first-order gradients themselves does

not exist. (Only for second-order gradients do we have something similar:

$$T_{zz} = -(T_{xx} + T_{yy}) .)$$

We add (22-16) and (22-17) and substitute the sum into (22-19), obtaining

$$\frac{\partial^2 K}{\partial z \partial z'} = -\frac{1}{\rho} \frac{\partial F}{\partial \rho} - \frac{\partial^2 F}{\partial \rho^2} .$$

In the plane $z = 0$ we have with (22-5)

$$\frac{\partial^2 K}{\partial z \partial z'} = -\frac{1}{\rho} K'(\rho) - K''(\rho) . \quad (22-20)$$

This relation expresses the covariance function of the *vertical* gradient T_z in terms of *horizontal* derivatives of the basic covariance function $K(\rho)$.

Let us similarly compute the covariance functions of the horizontal gradients T_x and T_y . We put

$$\cos \alpha = \frac{x' - x}{\rho} , \quad \sin \alpha = \frac{y' - y}{\rho} ; \quad (22-21)$$

α is thus the azimuth of the line PQ. Then (22-16) and (22-17) reduce for $z = z' = 0$ to

$$\frac{\partial^2 K}{\partial x \partial x'} = K_1(\rho) \cos^2 \alpha + K_t(\rho) \sin^2 \alpha , \quad (22-22)$$

$$\frac{\partial^2 K}{\partial y \partial y'} = K_1(\rho) \sin^2 \alpha + K_t(\rho) \cos^2 \alpha ,$$

where

$$K_1(\rho) = -K''(\rho) \quad (22-23)$$

represents the *longitudinal covariance*, and

$$K_t(\rho) = -\frac{1}{\rho} K'(\rho) \quad (22-24)$$

is the *transversal covariance*; cf. (Grafarend, 1976) and (Moritz, 1973a, pp. 64-66).

These expressions connect also the autocovariance functions of gravity anomalies and of vertical deflections, since Δg , ξ , η are related to T_x , T_y , T_z in a very simple way; cf. (15-8) and (15-9). This may be useful when determining covariance functions from data of different type.

All these relations hold in the planar approximation, that is, in a local area.

Essential parameters of covariance functions. There arises the question whether covariance functions can be satisfactorily characterized by means of only a few parameters. This is possible as regards the local behavior of the covariance function of gravity anomalies: in this case there are three "essential" parameters, for which we may take the variance C_0 , the correlation length ξ , and the curvature parameter χ (or, alternatively, the gradient variance G_0). These parameters are defined as follows.

Consider the covariance function $C(\psi)$ at sea level, which is the function (22-2) for $r = r' = R$. In the planar approximation this is a function $C(\rho)$ of the distance (22-4), which is related to the spherical distance ψ by

$$\rho = R\psi . \quad (22-25)$$

Clearly $C(\rho)$ is completely analogous to $K(\rho)$ as given by (22-5), the difference being that $C(\rho)$ is the covariance function for Δg and $K(\rho)$ denotes the covariance function for T . Fig. 22.1 shows the graph of such a function $C(\rho)$.

Now the *variance* C_0 is the value of the covariance function $C(\rho)$ for $\rho = 0$:

$$C_0 = C(0) . \quad (22-26)$$

The *correlation length* (in German: Halbwertsbreite) ξ is the value of the argument for which $C(\rho)$ has decreased to half of its value at $\rho = 0$:

$$C(\xi) = \frac{1}{2} C_0 . \quad (22-27)$$

The *curvature parameter* χ is a dimensionless quantity related to the curvature κ of the covariance curve at $\rho = 0$ by

$$\chi = \kappa \xi^2 / C_0 . \quad (22-28)$$

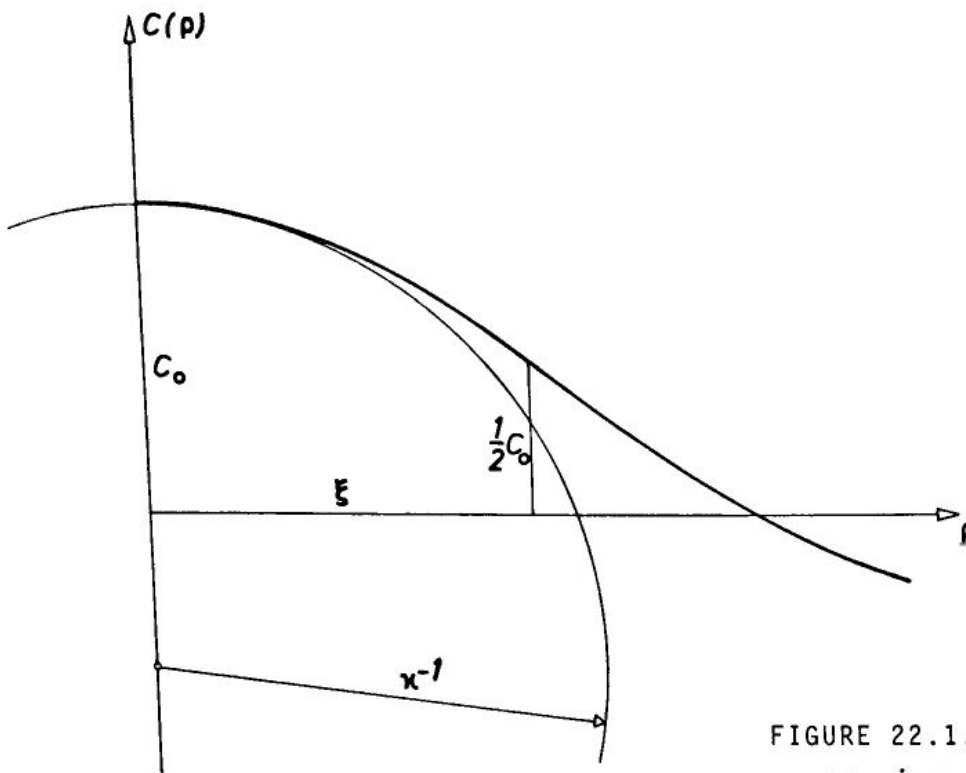


FIGURE 22.1. Local parameters of a covariance function.

The well-known formula for the curvature of the curve $C(\rho)$ gives

$$\kappa = \frac{C''}{(1+C'^2)^{3/2}}, \quad (22-29)$$

where

$$C' = \frac{dC}{d\rho}, \quad C'' = \frac{d^2C}{d\rho^2}.$$

For $\rho = 0$ we have $C'(0) = 0$ since the tangent to the curve is horizontal at the origin, and we get

$$\kappa = -C''(0). \quad (22-30)$$

The minus sign is conventional; it corresponds to taking the negative square root in (22-29).

The function $C(P, Q)$ is related to Δg in the same way as $K(P, Q)$ is related to T . Therefore,

$$C_1(\rho) = -C''(\rho) , \quad C_t(\rho) = -\frac{1}{\rho} C'(\rho) \quad (22-31)$$

represent the longitudinal and transversal covariance functions of horizontal anomalous gravity gradients $\partial\Delta g/\partial x$ and $\partial\Delta g/\partial y$, in analogy to (22-23) and (22-24), and

$$\frac{\partial^2 C}{\partial z \partial z^*} = -\frac{1}{\rho} C'(\rho) - C''(\rho) \quad (22-32)$$

gives the covariance function of the vertical anomalous gravity gradient $\partial\Delta g/\partial z$, in analogy to (22-20).

It should be mentioned that, as a planar approximation, the function $C(P,Q)$ is harmonic together with $K(P,Q)$. In fact, for $R \rightarrow \infty$, eq.(11-1) reduces to

$$\Delta g = -\frac{\partial T}{\partial z} .$$

Hence to this approximation,

$$\Delta(\Delta g) = \Delta\left(-\frac{\partial T}{\partial z}\right) = -\frac{\partial}{\partial z}(\Delta T) = 0$$

when $\Delta T = 0$, so that Δg is harmonic if T is (it is clear that the symbol Δ denotes the Laplace operator except in Δg). The harmonicity of $C(P,Q)$ follows from that of Δg , and therefore the relation (22-32) holds, which presupposes that $C(P,Q)$ is harmonic.

At the origin, for $\rho = 0$, eq. (22-31) gives

$$C_1(0) = -C''(0) = G_0$$

which defines the *gradient variance* G_0 . By de l'Hopital's rule, also

$$C_t(0) = -\lim_{\rho \rightarrow 0} \frac{C'(\rho)}{\rho} = -\lim_{\rho \rightarrow 0} \frac{C''(\rho)}{1} = -C''(0) = G_0 .$$

Hence, by (22-22),

$$\left(\frac{\partial^2 C}{\partial x \partial x^*}\right)_{\rho=0} = \left(\frac{\partial^2 C}{\partial y \partial y^*}\right)_{\rho=0} = G_0 , \quad (22-33)$$

so that G_0 is the variance of any horizontal gradient. Eq. (22-32) reduces for $\rho = 0$ to

$$\left(\frac{\partial^2 C}{\partial z \partial z} \right)_{\rho=0} = -2C''(0) = 2G_0, \quad (22-34)$$

so that the variance of the vertical gradient is $2G_0$.

The comparison between (22-30) and (22-34) gives

$$\kappa = G_0, \quad (22-35)$$

so that (22-28) may be written as

$$\chi = \xi^2 G_0 / C_0. \quad (22-36)$$

Thus the gradient variance G_0 and the curvature parameter χ are related in a simple way and are essentially equivalent.

The definition of the gradient variance G_0 should be carefully noted. It is either the variance of any anomalous horizontal gravity gradient or, equivalently, half of the variance of the anomalous vertical gravity gradient:

$$\begin{aligned} G_0 &= \text{cov} \left(\frac{\partial \Delta g}{\partial x}, \frac{\partial \Delta g}{\partial x} \right)_{\rho=0} = \text{var} \left(\frac{\partial \Delta g}{\partial x} \right) = \text{var} \left(\frac{\partial \Delta g}{\partial y} \right) = \\ &= \frac{1}{2} \text{cov} \left(\frac{\partial \Delta g}{\partial z}, \frac{\partial \Delta g}{\partial z} \right)_{\rho=0} = \frac{1}{2} \text{var} \left(\frac{\partial \Delta g}{\partial z} \right); \end{aligned} \quad (22-37)$$

cf. (22-33) and (22-34).

Test computations with various functions $C(\rho)$, all having the same C_0 , ξ and χ , show that the functional values $C(\rho)$ for $\rho \leq \xi$ are practically the same for all "reasonable" analytical expressions for $C(\rho)$; differences occur only for $\rho > \xi$ (Moritz, 1976b, p.31). Now the domain $0 \leq \rho \leq \xi$ is the one most important for interpolation: interpolation is accurate enough only if station distances are well below ξ . In this sense we may say that all functions $C(\rho)$ having the same parameters C_0 , ξ , χ are practically equivalent for many applications.

Positive definiteness. As we have seen in sec.9, all covariance matrices must be positive definite. A function $K(P, Q)$ of type (10-9) is called positive definite if all coefficients k_n are nonnegative (positive or

zero). It can then be shown that all regular signal covariance matrices C_{tt} derived from such a function by covariance propagation are positive definite matrices, cf. sec. 24.

Consider the function K restricted to the sphere, as given by (10-7). A spherical-harmonic expansion of a function defined on the sphere is also called a spectral representation, thus the coefficients k_n in (10-7) form the spectrum of the covariance function $K(\psi)$.

Alternatively we may consider the covariance function $C(P, Q)$ of the gravity anomaly (22-2), which for $r = r' = R$ reduces to

$$C(\psi) = \sum_{n=2}^{\infty} c_n P_n(\cos \psi) . \quad (22-38)$$

All c_n are nonnegative by (22-3), so that $C(\psi)$ is also positive definite as it should be.

In the planar approximation, where the basic sphere $r = R$ is replaced by the xy-plane $z = 0$, the corresponding spectral representation is given by

$$C(\rho) = \int_0^{\infty} \bar{C}(n) J_0(n\rho) n dn , \quad (22-39)$$

where $J_0(x)$ is the Bessel function of zero order, and $\bar{C}(n)$ is the spectrum of the function $C(\rho)$. The inverse formula has the same structure:

$$\bar{C}(n) = \int_0^{\infty} C(\rho) J_0(n\rho) \rho d\rho , \quad n \geq 0 . \quad (22-40)$$

The formulas (22-39) and (22-40) define a *Hankel transformation*. It is related to the well-known Fourier transformation: the two-dimensional Fourier transform for isotropic functions (which depend only on ρ but not on the azimuth α) reduces to the Hankel transform (Papoulis, 1968, p.140).

Positive definiteness of a function $C(\rho)$ is again equivalent to the *nonnegativity of the spectrum*: in the present case this means that the Hankel transform of $C(\rho)$ is everywhere positive or zero:

$$\bar{C}(n) \geq 0 \quad \text{if} \quad n \geq 0 . \quad (22-41)$$

Examples of plane covariance functions. The best-known positive definite function in the plane is the Gaussian function

$$C_1(\rho) = C_0 e^{-A^2 \rho^2}, \quad (22-42)$$

C_0 and A being two constants. It has the unique property that its Hankel transform is also a Gaussian function:

$$\bar{C}_1(n) = \frac{C_0}{2A^2} e^{-n^2/4A^2}, \quad (22-43)$$

which obviously is everywhere positive (Papoulis, 1968, p.145).

The basic parameters for this function $C_1(\rho)$ are readily found. The variance is the constant already denoted by C_0 in (22-42). The correlation length is obtained from (22-27) to be

$$\xi_1 = \frac{1}{A} \sqrt{\ln 2}, \quad (22-44)$$

and the curvature parameter is found from (22-28) and (22-30):

$$\chi_1 = 2 \ln 2 = 1.386..; \quad (22-45)$$

the symbol " \ln " denotes the natural logarithm.

Unfortunately, the Gaussian function (22-43) does not have a simple harmonic extension into outer space $z > 0$, so that it cannot very well be used as a spatial covariance function.

Therefore we may consider other analytical models, for instance

$$C(\rho) = \frac{C_0}{(1 + B^2 \rho^2)^m}, \quad (22-46)$$

with constants C_0 , B , and m . An example, for $m = 2$, is Hirvonen's covariance function (Heiskanen and Moritz, 1967, p.255).

It can be shown (Moritz, 1976b, p.42) that the function (22-46) admits a simple harmonic extension into the region $z > 0$ only for $m = 1/2$ and $m = 3/2$. We then obtain the spatial harmonic functions

$$C_2(P, Q) = \frac{C_0 b}{[\rho^2 + (z + z' + b)^2]^{1/2}}, \quad (22-47)$$

$$C_3(P, Q) = \frac{C_0 b^2 (z+z'+b)}{[\rho^2 + (z+z'+b)^2]^{3/2}} \quad (22-48)$$

Here b is a constant, ρ is again given by (22-4), and P and Q have the coordinates (x, y, z) and (x', y', z') , respectively. For $z = z' = 0$, these functions reduce to (22-46), with $B = b^{-1}$:

$$C_2(\rho) = \frac{C_0}{(1+B^2 \rho^2)^{1/2}} \quad (22-49)$$

$$C_3(\rho) = \frac{C_0}{(1+B^2 \rho^2)^{3/2}} \quad (22-50)$$

The spectrum, or Hankel transform, of these two functions is simple (Papoulis, 1968, p.145): we have

$$\bar{C}_2(n) = \frac{C_0}{B} \frac{e^{-n/B}}{n} \quad (22-51)$$

$$\bar{C}_3(n) = \frac{C_0}{B^2} e^{-n/B} \quad (22-52)$$

These spectra are certainly positive for $n > 0$, so that the functions $C_2(\rho)$ and $C_3(\rho)$ are also positive definite.

Of the three basic parameters of these functions, the variance has already been denoted by C_0 , and for the other two parameters we readily find

$$\xi_2 = \sqrt{3}/B, \quad x_2 = 3 \quad (22-53)$$

for the function $C_2(\rho)$, and

$$\xi_3 = (2^{2/3}-1)^{1/2}/B, \quad x_3 = 3(2^{2/3}-1) = 1.762.. \quad (22-54)$$

for the function $C_3(\rho)$.

We finally remark that in choosing an analytical approximation to an empirically determined covariance function, we must be careful to use a positive definite function. For instance, a polynomial

$$p(\rho) = a_0 + a_1 \rho^2 + a_2 \rho^4 + \dots a_n \rho^{2n} \quad (22-55)$$

is not positive definite and cannot be used as an approximation for $C(\rho)$.

23. GLOBAL COVARIANCE MODELS

We shall start with the consideration of the covariance function of the gravity anomaly as represented by (22-2):

$$C(P, Q) = \sum_{n=0}^{\infty} c_n \left(\frac{R^2}{rr'} \right)^{n+2} P_n(\cos \psi) ; \quad (23-1)$$

we have begun the sum with $n=0$. By putting

$$c_n = \gamma_n \left(\frac{R_B}{R} \right)^{2n+4} \quad (23-2)$$

we may transform this basic expression into the form

$$C(P, Q) = \sum_{n=0}^{\infty} \gamma_n \left(\frac{R_B^2}{rr'} \right)^{n+2} P_n(\cos \psi) , \quad (23-3)$$

where $R_B < R$ is the radius of a sphere concentric to the terrestrial sphere and of slightly smaller radius; the sphere of radius R_B is frequently called a "Bjerhammar sphere" (see also p.69).

By means of the further substitution

$$\sigma = \frac{R_B^2}{rr'} , \quad (23-4)$$

the covariance function may simply be written as

$$C(P, Q) = \sum_{n=0}^{\infty} \gamma_n \sigma^{n+2} P_n(\cos \psi) . \quad (23-5)$$

Basic models. We shall now consider three simple cases in which this series can be summed in closed form.

The first case is

$$\gamma_n = 1 \quad (23-6)$$

for all n , so that

$$C(P, Q) = \sum_{n=0}^{\infty} \sigma^{n+2} P_n(\cos \psi) . \quad (23-7)$$

This series may be summed by means of the well-known expression for the reciprocal distance; cf. eq. (1-80) of (Heiskanen and Moritz, 1967, p.33):

$$\frac{1}{\sqrt{1-2\sigma t+\sigma^2}} = \sum_{n=0}^{\infty} \sigma^n P_n(t) . \quad (23-8)$$

With

$$t = \cos \psi \quad (23-9)$$

and

$$L = \sqrt{1-2\sigma \cos \psi + \sigma^2} , \quad (23-10)$$

eq. (23-7) thus reduces to

$$C(P, Q) = \frac{\sigma^2}{L} . \quad (23-11)$$

This function may be called the *reciprocal distance covariance function*, although L is not simply the spatial distance between the points P and Q .

The local behavior of this function may be studied by considering its planar approximation. To perform a suitable transition from the sphere to its tangent plane we put

$$\zeta = 1 - \sigma , \quad (23-12)$$

$$\lambda = 2 \sin \frac{\psi}{2} . \quad (23-13)$$

Since

$$\cos \psi = 1 - 2 \sin^2 \frac{\psi}{2} = 1 - \frac{1}{2} \lambda^2, \quad (23-14)$$

eq. (23-10) is easily found to reduce to

$$L^2 = \zeta^2 + \sigma \lambda^2; \quad (23-15)$$

this equation is still rigorous.

We now put

$$\begin{aligned} r &= R + z, \\ r' &= R + z', \\ R_B &= R - b/2. \end{aligned} \quad (23-16)$$

Then (23-4) gives

$$\sigma = \frac{(R-b/2)^2}{(R+z)(R+z')} = 1 - \frac{z+z'+b}{R} + \dots \quad (23-17)$$

From (23-12) we thus get

$$\zeta = \frac{z+z'+b}{R} [1 + O(\zeta)], \quad (23-18)$$

where $O(\zeta)$ denotes terms of order ζ or smaller. Similarly, (23-13) gives

$$\lambda = \frac{\rho}{R} [1 + O(\lambda^2)], \quad (23-19)$$

where ρ is the horizontal distance (22-4). Also (23-12) may be written in an analogous way:

$$\sigma = 1 + O(\zeta). \quad (23-20)$$

Thus (23-15) gives

$$L = \frac{1}{R} \sqrt{\rho^2 + (z+z'+b)^2}, \quad (23-21)$$

disregarding relative errors $O(\zeta)$ and $O(\lambda^2)$. On admitting a constant factor we thus see that (23-11), as a planar approximation, reduces to the plane covariance function (22-47).

Since the curvature parameter χ expresses a strictly local property, it has the same value for a spherical covariance function and for its planar approximation. Thus, by (22-53), the curvature parameter for the reciprocal distance covariance function is

$$\chi = 3. \quad (23-22)$$

Our second basic model is obtained by putting

$$\gamma_n = 2n+1, \quad (23-23)$$

so that (23-5) becomes

$$C(P,Q) = \sum_{n=0}^{\infty} (2n+1) \sigma^{n+2} P_n(\cos \psi). \quad (23-24)$$

By differentiating the basic identity (23-8) with respect to σ , multiplying by $2\sigma^3$, and adding σ^2/L we get

$$2\sigma^3 \frac{\partial}{\partial \sigma} \left(\frac{1}{L} \right) + \frac{\sigma^2}{L} = \sum_{n=0}^{\infty} (2n+1) \sigma^{n+2} P_n(\cos \psi). \quad (23-25)$$

The right-hand side is equal to (23-24), and the left-hand side can be computed by explicitly differentiating $1/L$ as a function of σ , as given by (23-10). Thus (23-24) becomes

$$C(P,Q) = \frac{\sigma^2(1-\sigma^2)}{L^3}, \quad (23-26)$$

which is the desired closed expression for the series (23-24).

The function (23-26) may be called the *Poisson covariance function*, because essentially it represents the kernel in the well-known Poisson integral (cf. Heiskanen and Moritz, 1967, p.35). It has been given by Krarup (1969, p.43) and has also been used, e.g., by Jordan (1978).

The planar approximation to this function is readily seen to be (22-48), so that the curvature parameter is given by (22-54):

$$\chi = 1.762 . \quad (23-27)$$

As a third basic model we consider the case

$$\gamma_n = \frac{1}{n} \quad (n \geq 1) \quad (23-28)$$

giving

$$C(P, Q) = \sum_{n=1}^{\infty} \frac{1}{n} \sigma^{n+2} P_n(\cos \psi) . \quad (23-29)$$

To sum this series we form

$$\frac{1}{\sigma^2} C(P, Q) = \sum_{n=1}^{\infty} \frac{1}{n} \sigma^n P_n(\cos \psi)$$

and differentiate with respect to σ , obtaining

$$\frac{\partial}{\partial \sigma} \left(\frac{C}{\sigma^2} \right) = \sum_{n=1}^{\infty} \sigma^{n-1} P_n(\cos \psi) .$$

The comparison with (23-8) gives, in view of (23-10),

$$\frac{\partial}{\partial \sigma} \left(\frac{C}{\sigma^2} \right) = \frac{1}{\sigma} \left(\frac{1}{L} - 1 \right) ,$$

so that

$$\sigma^{-2} C(P, Q) = \int \frac{1}{\sigma L} d\sigma - \int \frac{d\sigma}{\sigma} + k ; \quad (23-30)$$

the integration constant k is to be determined from the condition $\sigma^{-2} C = 0$ for $\sigma = 0$, which follows from (23-29).

The second integral is, of course, $-\ln \sigma$, but also the first integral is standard, so that it can be found in an integral table such as (Gradshteyn and Ryzhik, 1965). Using no. 2.266 of this table, p.84, we thus readily obtain

$$C(P, Q) = \sigma^2 \ln \frac{2}{N} , \quad (23-31)$$

where we have put

$$N = 1 + L - \sigma \cos \psi . \quad (23-32)$$

The function (23-31) may be called a *logarithmic covariance function*.

It is found that the basic constants of this function, correlation length ξ and curvature parameter χ , are approximately given by

$$\xi = 2R\sqrt{\zeta} , \quad \chi = \frac{2}{\zeta \ln \zeta - 1} , \quad (23-33)$$

ζ being defined by (23-12) (Moritz, 1976b, pp.47-48). As an example, for $\xi = R/100 = 63.7$ km we get

$$\chi = 7550 , \quad (23-34)$$

which is quite large as compared to (23-22) or (23-27).

Auxiliary formulas. For later use we shall sum the following series, first considered by Tscherning (1972):

$$F_A(\sigma, t) = \sum_{n=0}^{\infty} \frac{1}{n+A} \sigma^{n+1} p_n(t) \quad \text{for } A > 0 , \quad (23-35)$$

$$F_A(\sigma, t) = \sum_{n=1-A}^{\infty} \frac{1}{n+A} \sigma^{n+1} p_n(t) \quad \text{for } A \leq 0 , \quad (23-36)$$

where A denotes some fixed integer; the summations are over n . For $A = 0$, $F_0(\sigma, t)$ as given by (23-36) differs from the logarithmic function (23-29) or (23-31) only by the factor σ . Also the other series can be summed in closed form, following the same procedure as used for (23-29).

For $A > 0$ we form, using (23-35),

$$\sigma^{A-1} F_A(\sigma, t) = \sum_{n=0}^{\infty} \frac{1}{n+A} \sigma^{n+A} p_n(t) ,$$

differentiate:

$$\frac{\partial}{\partial \sigma} (\sigma^{A-1} F_A) = \sum_{n=0}^{\infty} \sigma^{n+A-1} p_n(t) = \frac{\sigma^{A-1}}{L} , \quad (23-37)$$

and integrate:

$$\sigma^{A-1} F_A(\sigma, t) = \int_0^\sigma \frac{\sigma^{A-1}}{L} d\sigma. \quad (23-38)$$

The lower limit of the integral is zero since $F_A(\sigma, t) = 0$ if $\sigma = 0$ by (23-35).

For $A \leq 0$ we use (23-36), obtaining

$$\begin{aligned} \frac{\partial}{\partial \sigma} (\sigma^{A-1} F_A) &= \sum_{n=0}^{\infty} \sigma^{n+A-1} p_n(t) \\ &= \frac{\sigma^{A-1}}{L} - \sum_{n=0}^{-1-A} \sigma^{n+A-1} p_n(t) - \frac{1}{\sigma} p_{-A}(t), \end{aligned} \quad (23-39)$$

whence

$$\sigma^{A-1} F_A = \int \frac{\sigma^{A-1}}{L} d\sigma - \sum_{n=0}^{-1-A} \frac{\sigma^{n+A}}{n+A} p_n(t) - \ln \sigma \cdot p_{-A}(t) + k_A; \quad (23-40)$$

the integration constant k_A is to be determined by the condition that $\sigma^{A-1} F_A(\sigma, t) = 0$ if $\sigma = 0$.

The integrals (23-38) and (23-40) can be evaluated by the help of the integral table (Gradshteyn and Ryzhik, 1965).

Using no. 2.269, *ibid.*, p.85, we get

$$F_0(\sigma, \psi) = \sigma \ln \frac{2}{N}, \quad (23-41)$$

$$F_{-1}(\sigma, \psi) = \sigma(M + \sigma t \cdot \ln \frac{2}{N}), \quad (23-42)$$

$$F_{-2}(\sigma, \psi) = \frac{1}{2} \sigma(1+3\sigma t)M + \sigma^3 p_2(t) \ln \frac{2}{N} + \frac{1}{4} \sigma^3 (1-t^2), \quad (23-43)$$

where

$$M = 1 - L - \sigma t, \quad (23-44)$$

$$N = 1 + L - \sigma t, \quad (23-45)$$

L being defined by (23-10) as usual and $t = \cos \psi$; as a matter of fact, (23-41) is equivalent to (23-31) since (23-29) differs from F_0 only by the factor σ ; see also (Tscherning and Rapp, 1974, pp.32-35).

Other negative A will not be used.

For positive A we get from no. 2.264, *ibid.*, p.83:

$$F_1(\sigma, \psi) = \ln \left(1 + \frac{2\sigma}{1-\sigma+L} \right), \quad (23-46)$$

$$F_2(\sigma, \psi) = \sigma^{-1} [L - 1 + t F_1(\sigma, \psi)] . \quad (23-47)$$

Functions F_A for larger positive A can be found by a recursion formula which is obtained in the following way. The differentiation of (23-10) gives

$$\frac{\partial L}{\partial \sigma} = \frac{\sigma - t}{L}, \quad (23-48)$$

so that

$$\frac{\partial}{\partial \sigma} (\sigma^{A-1} L) = (A-1) \sigma^{A-2} L + \sigma^{A-1} \frac{\sigma - t}{L},$$

and after some straightforward algebra,

$$\frac{\partial}{\partial \sigma} (\sigma^{A-1} L) = A \frac{\sigma^A}{L} - (2A-1)t \frac{\sigma^{A-1}}{L} + (A-1) \frac{\sigma^{A-2}}{L} .$$

We integrate with respect to σ and take (23-38) into account, obtaining

$$\sigma^{A-1} L = A \sigma^A F_{A+1} - (2A-1)t \sigma^{A-1} F_A + (A-1) \sigma^{A-2} F_{A-1} ,$$

whence

$$F_{A+1} = \frac{1}{A\sigma} \left[L + (2A-1)t F_A - (A-1) \sigma^{-1} F_{A-1} \right] . \quad (23-49)$$

This is the desired recursion formula.

Besides the F_A we shall also use the function

$$F(\sigma, \psi) = \sum_{n=0}^{\infty} \sigma^{n+1} P_n(t) = \frac{\sigma}{L}, \quad (23-50)$$

which differs from the reciprocal distance covariance function (23-11) only by the factor σ .

The data. The global variance C_0 of the gravity anomaly has been determined on the basis of a large amount of gravity data distributed over the whole earth by Tscherning and Rapp (1974); they obtain

$$C_0 = 1795 \text{ mgal}^2 \quad (23-51)$$

The other local parameters ξ and G_0 are far less reliably known. For the correlation length ξ , defined by (22-27), we have estimates ranging from about 40 km to 80 km; cf. (Schwarz, 1976b, p.14).

Especially poorly determined is the gradient variance G_0 , defined by (22-37) as the variance of any anomalous horizontal gravity gradient or, equivalently, as half of the variance of the anomalous vertical gradient. The literature seems to favor values for G_0 on the order of $2 (\text{mgal/km})^2$, but these values are little more than rough guesses (Schwarz, 1976b, p.15; Tscherning, 1976, p.38; Moritz, 1977a, p.2; Schwarz, 1978a, p.103).

In fact, the gravity gradients are very irregular and highly sensitive to local perturbing masses. Every geodetic application of Δg and of its gradients involves--explicitly or implicitly--some smoothing, depending on how much local detail we are willing to take into consideration. For instance, any use of mean values of Δg , even if the block size is as small as $1' \times 1'$, implies such a smoothing. In view of their very irregularity, gravity gradients are particularly strongly influenced by such a smoothing.

Therefore, a theoretically valid and practically useful determination of G_0 does not only depend on the gradient data: we must also carefully define the way in which we wish to smooth our data.¹

On the other hand, a precise definition of the correlation length ξ depends on whether we wish to consider a local or a truly global covariance function; cf. the corresponding remark at the end of this section.

A comprehensive study on these questions from a practical point of view would be highly desirable.

Besides the "local" parameters C_0 , ξ , and G_0 (so called because they describe the local behavior of $C(P,Q)$ for small distances PQ), we have "global" parameters characterizing the function $C(P,Q)$ as a whole, namely the degree variances c_n for lower degree n (say, $n \leq 20$). These c_n can be obtained from satellite observations, preferably combined with gravimetry.

¹ Generally we suggest that an exact specification of the degree of smoothing would be essential to any geodetic application of gravimetric methods, for instance, to the practical solution of boundary-value problems.

Table 23.1 shows two sets of such c_n , $3 \leq n \leq 20$.

n	Rapp 1973	GEM 10	n	Rapp 1973	GEM 10
3	33.9	33.5	12	4.8	3.6
4	19.2	19.6	13	11.7	6.2
5	21.6	20.6	14	5.5	3.4
6	18.9	19.0	15	7.3	3.0
7	18.8	19.1	16	6.5	2.6
8	10.4	11.4	17	5.7	2.1
9	11.1	11.1	18	10.7	3.1
10	11.4	9.7	19	11.0	2.8
11	8.4	6.6	20	8.9	2.0

TABLE 23.1. Degree variances c_n .

The first set is from (Rapp, 1973); it is given because the covariance function of (Tscherning and Rapp, 1974) is based on it. The second set is for Goddard Earth Model (GEM) 10 (Lerch et al., 1977).

Covariance model fitting. The local parameters C_0 , ξ , G_0 and the global parameters c_3 through c_{20} (say) are the basic data to which an analytical expression for the covariance function is to be fitted.

The Poisson covariance function (23-26) appears to be excluded for global purposes by the fact that the degree variances (which are essentially γ_n by (23-2)) increase with n by (23-23), instead of decreasing according to Table 23.1. Furthermore, the curvature parameter (23-27) seems to be too small.

Tscherning and Rapp (1974) have modeled the general trend of the degree variances c_3 to c_{20} , as well as the variance C_0 , by means of the expression

$$C(P, Q) = \alpha \sum_{n=3}^{\infty} \frac{n-1}{(n-2)(n+B)} \sigma_0^{n+2} \left(\frac{R^2}{rr'} \right)^{n+2} P_n(\cos \psi). \quad (23-52)$$

Here σ_0 is the constant value of σ , as given by (23-4), for $r = r' = R$:

$$\sigma_0 = \frac{R_B^2}{R^2}. \quad (23-53)$$

Now the local behavior of a function is mainly influenced by the spherical harmonics of high degree n , the lower degrees corresponding to long waves which are almost constant in small regions. But for large n we have

$$\frac{n-1}{(n-2)(n+B)} = \frac{1 - \frac{1}{n}}{n \left(1 - \frac{2}{n}\right) \left(1 + \frac{B}{n}\right)} \doteq \frac{1}{n}, \quad (23-54)$$

so that locally the function (23-52) behaves practically as a logarithmic covariance function (23-29). Owing to the integer B , however, the expression (23-52) has a greater flexibility for fitting.

By fitting the variance (23-51) and the degree variances shown in Table 23.1, first column, Tscherning and Rapp (1974, p.22) get

$$\begin{aligned} \alpha &= 425.28 \text{ mgal}^2, \\ B &= 24, \\ \sigma_0 &= 0.999617. \end{aligned} \quad (23-55)$$

The correlation length ξ and the gradient variance G_0 are

$$\xi = 42 \text{ km}, \quad G_0 = 35.4 (\text{mgal/km})^2. \quad (23-56)$$

By what has been said above, the value for G_0 appears considerably too large; the corresponding curvature parameter is

$$\chi = 34.8. \quad (23-57)$$

This is due to the fact that the Tscherning-Rapp covariance function is essentially a logarithmic covariance function which is characterized by large χ ; cf. (23-34). It is, therefore, natural to try a linear combination of (23-52) with a function (23-11) which has a χ of only 3.

We are thus led to the model

$$\begin{aligned} C(P, Q) &= \alpha_1 \sum_3^\infty \frac{n-1}{n+A} \sigma_1^{n+2} \left(\frac{R^2}{rr'} \right)^{n+2} P_n(\cos\psi) + \\ &+ \alpha_2 \sum_3^\infty \frac{n-1}{(n-2)(n+B)} \sigma_2^{n+2} \left(\frac{R^2}{rr'} \right)^{n+2} P_n(\cos\psi), \end{aligned} \quad (23-58)$$

which contains six free parameters: $\alpha_1, \alpha_2, \sigma_1, \sigma_2$, and the integers A and B . Here σ_1 and σ_2 are related to two different values R_1 and R_2 for R_B by

$$\sigma_1 = \frac{R_1^2}{R^2}, \quad \sigma_2 = \frac{R_2^2}{R^2}, \quad (23-59)$$

corresponding to (23-53).

This is the covariance function for the gravity anomaly Δg . The corresponding covariance function for the anomalous potential T is given by

$$\begin{aligned} K(P, Q) = & \alpha_1 \sum_{n=3}^{\infty} \frac{R_1^2}{(n-1)(n+A)} \sigma_1^{n+1} \left(\frac{R^2}{rr'} \right)^{n+1} P_n(\cos \psi) \\ & + \alpha_2 \sum_{n=3}^{\infty} \frac{R_2^2}{(n-1)(n-2)(n+B)} \sigma_2^{n+1} \left(\frac{R^2}{rr'} \right)^{n+1} P_n(\cos \psi). \end{aligned} \quad (23-60)$$

If $A \neq -1$, $B \neq -1$ and $\neq -2$, then all these series can be summed in closed form by a decomposition into partial fractions, using the functions F_A and F introduced above. We have

$$\frac{n-1}{n+A} = 1 - \frac{A+1}{n+A}, \quad (23-61)$$

$$\frac{n-1}{(n-2)(n+B)} = \frac{1}{B+2} \left(\frac{1}{n-2} + \frac{B+1}{n+B} \right), \quad (23-62)$$

$$\frac{1}{(n-1)(n+A)} = \frac{1}{A+1} \left(\frac{1}{n-1} - \frac{1}{n+A} \right), \quad (23-63)$$

$$\frac{1}{(n-1)(n-2)(n+B)} = \frac{1}{(B+1)(B+2)} \left(\frac{B+1}{n-2} - \frac{B+2}{n-1} + \frac{1}{n+B} \right). \quad (23-64)$$

Let us write (23-58) and (23-60) in the form

$$C(P, Q) = \alpha_1 C_1(P, Q) + \alpha_2 C_2(P, Q), \quad (23-65)$$

$$K(P, Q) = \alpha_1 K_1(P, Q) + \alpha_2 K_2(P, Q).$$

Then we readily find for positive nonzero integers A and B :

$$C_1(P, Q) = \sigma_1 \left[F(\sigma_1, \psi) - \sigma_1 - \frac{\sigma_1^2 t}{A} - \frac{\sigma_1^3}{A+2} P_2(t) \right] - \\ - (A+1) \sigma_1 \left[F_A(\sigma_1, \psi) - \frac{\sigma_1}{A} - \frac{\sigma_1^2 t}{A+1} - \frac{\sigma_1^3}{A+2} P_2(t) \right], \quad (23-66)$$

$$C_2(P, Q) = \frac{1}{B+2} \sigma_2 F_{-2}(\sigma_2, \psi) + \\ + \frac{B+1}{B+2} \sigma_2 \left[F_B(\sigma_2, \psi) - \frac{\sigma_2}{B} - \frac{\sigma_2^2 t}{B+1} - \frac{\sigma_2^3}{B+2} P_2(t) \right]; \quad (23-67)$$

$$K_1(P, Q) = \frac{R_1^2}{A+1} \left[F_{-1}(\sigma_1, \psi) - \sigma_1^3 P_2(t) \right] - \\ - \frac{R_1^2}{A+1} \left[F_A(\sigma_1, \psi) - \frac{\sigma_1}{A} - \frac{\sigma_1^2 t}{A+1} - \frac{\sigma_1^3}{A+2} P_2(t) \right], \quad (23-68)$$

$$K_2(P, Q) = \frac{R_2^2}{B+2} F_{-2}(\sigma_2, \psi) - \\ - \frac{R_2^2}{B+1} \left[F_{-1}(\sigma_2, \psi) - \sigma_2^3 P_2(t) \right] + \\ + \frac{R_2^2}{(B+1)(B+2)} \left[F_B(\sigma_2, \psi) - \frac{\sigma_2}{B} - \frac{\sigma_2^2 t}{B+1} - \frac{\sigma_2^3}{B+2} P_2(t) \right]. \quad (23-69)$$

To repeat, the integers A and B are > 0 in these formulas. The respective functions F_A and F_B are therefore given by the recursion formula (23-49) together with (23-46) and (23-47). The functions F_{-1} and F_{-2} are, of course, expressed by (23-42) and (23-43), and F is (23-50).

Values of A and $B \leq -3$ are impossible because otherwise a denominator in the series (23-58) or (23-60) would be zero. The values $B = -1$ and -2 are excluded because then these series cannot be summed. This eliminates all negative values for B ; for A , only the negative value -2 is possible.

The formulas (23-66) through (23-69) hold only for A and $B > 0$ and cannot be directly used if A or B is zero. It is easily seen, however, that they remain valid even for $A = 0$ or $B = 0$ or both, provided the terms σ_1/A and σ_2/B are omitted whenever A or B are zero. We can use (23-66) and (23-68) even for $A = -2$ provided we replace the second bracket in these formulas simply by $F_{-2}(\sigma_1, \psi)$, omitting the following terms (Moritz, 1977a, pp.7-8).

We also mention that the gradient variance G_0 is given, to a sufficient approximation, by

$$G_0 = \frac{\alpha_1}{R^2 \xi_1^3} + \frac{\alpha_2}{2R^2 \xi_2^2} , \quad (23-70)$$

where

$$\xi_1 = 1 - \sigma_1 , \quad \xi_2 = 1 - \sigma_2 \quad (23-71)$$

(*ibid.*, p.18).

Various attempts to fit the models (23-52) and (23-58) to available data are presented in (Jekeli, 1978).

For numerical computation of covariance functions of the type considered here, the Fortran subroutine COVAX developed by Tscherning (1976) can be used. It gives covariance functions for T as well as for first and second-order gradients. Computer time can be reduced by using COVAX for computing covariance function values at grid points and interpolating between them (Sünkel, 1978b).

Since lower-degree spherical harmonics obviously are almost constant locally, local or regional covariance functions may be obtained from global ones by subtracting all spherical harmonics for $3 < n < N$, for a suitable N (Tscherning and Rapp, 1974, p.62). This particularly affects the variance C_0 and also the correlation length ξ .

PART C

LEAST-SQUARES COLLOCATION: ADVANCED ASPECTS

A deeper theoretical understanding of the analytical structure of least-squares collocation involves the geometrical representation in a Hilbert space with a kernel function. The theory of such Hilbert spaces is reviewed in sec. 24, and sec. 25 presents the Hilbert space theory of collocation: mathematically, least-squares collocation can be regarded as a least-squares adjustment in an infinitely-dimensional Hilbert space.

In sections 26 to 30 we present a general treatment of the operational approach to physical geodesy. Starting with a given number of observational data, we ask how they can be used to determine the shape of the earth and its gravitational field. This problem, similar to inverse problems of geophysics, does not as such have a unique solution; it is a so-called improperly posed problem. Applying standard methods for such problems, we are lead to variational principles by means of which a unique solution can be achieved. In this way we again arrive at least-squares collocation; some alternatives are also pointed out.

Sections 30 to 38 contain a detailed study of statistical aspects of collocation. As an introduction we present a simple theory of stochastic processes on the circle. This permits a straightforward transition to the sphere: the anomalous gravity field may be considered a stochastic process on the sphere. It turns out that all relevant ergodic processes are non-Gaussian. The preferred model will use the theory of stochastic processes only as a formal mathematical apparatus but, physically, will work strictly with the one existing gravitational field only.

The last section (sec.39) presents an advanced aspect of a different character, the refinement of collocation because of very small ellipsoidal effects. So far, the theory of least-squares collocation was formally referred to a sphere. This is sufficient for most present practical purposes, but very precise computations may require ellipsoidal corrections.

24. HILBERT SPACES WITH KERNEL FUNCTIONS

A deeper mathematical understanding of least-squares collocation requires the theory of Hilbert spaces that possess a kernel function. Therefore we shall in this section present essential definitions and theorems for these spaces; further information can be found in the articles (Meissl, 1976), (Tscherning, 1978a), and in the book (Meschkowski, 1962).

Consider a Hilbert space H of functions $f(P)$, P being a point in a certain region B of three-dimensional space R^3 ; in our applications, B will be the exterior of a certain sphere of radius R_B . Remember that a Hilbert space is a complete inner product space satisfying (5-10).

Suppose now that there exists a function $K(P,Q)$ satisfying the two relations:

$$K(P,Q) \in H \text{ for } Q \text{ fixed,} \quad (24-1)$$

$$f(Q) = (f(P), K(P,Q))_P \text{ for all } f \in H. \quad (24-2)$$

The first relation says that $K(P,Q)$, considered as a function of P , belongs to the Hilbert space H . The second relation states that the inner product of f with K , both considered as functions of P -- hence the notation $(\cdot)_P$ --, reproduces f .

A function $K(P,Q)$ satisfying these two fundamental properties, is called a *reproducing kernel function* or, briefly, a *kernel function*. It can be shown that a given Hilbert space possesses at most one kernel function, so that the two conditions (24-1) and (24-2) determine $K(P,Q)$ uniquely.

It is easily proved that the kernel function is *symmetric and positive definite*.

In fact, applying (24-2) to the function $f(P) = K(P,R)$ we get

$$K(Q,R) = (K(P,R), K(P,Q))_P. \quad (24-3)$$

By the symmetry of the inner product this is equal to

$$(K(P,Q), K(P,R))_P = K(R,Q),$$

which shows the symmetry

$$K(Q,R) = K(R,Q).$$

(24-4)

The positive definiteness is proved as follows. Consider a finite linear combination

$$\sum_{k=1}^N \lambda_k K(P, P_k) ,$$

λ_k being arbitrary constants and P_k being N fixed points. Obviously, for all P ,

$$\begin{aligned} 0 &\leq \left\| \sum_{k=1}^N \lambda_k K(P, P_k) \right\|^2 = \left(\sum_k \lambda_k K(P, P_k), \sum_l \lambda_l K(P, P_l) \right)_P \\ &= \sum_{k,l} \lambda_k \lambda_l (K(P, P_k), K(P, P_l))_P . \end{aligned}$$

Using (24-3) we thus have

$$\sum_{k,l} \lambda_k \lambda_l K(P_k, P_l) \geq 0 , \quad (24-5)$$

which expresses the positive definiteness of the function $K(P, Q)$.

The condition (24-5) says that a function $K(P, Q)$ is positive definite if every matrix with elements

$$K(P_k, P_l) , \quad k, l = 1, 2, \dots, N , \quad (24-6)$$

is positive definite; cf. (9-26). Formerly, in secs. 12 and 22, we have used another definition: a function $K(P, Q)$ is positive definite if it has a positive spectrum. It may be shown that these two definitions are equivalent; see the corresponding remark at the end of the present section.

By putting $N = 1$, $\lambda_1 = 1$, $P_1 = P$ we get from (24-5):

$$K(P, P) \geq 0 . \quad (24-7)$$

The kernel function $K(P, Q)$ may be represented¹ by means of a complete orthonormal system $\phi_i(P)$:

$$K(P, Q) = \sum_{i=1}^{\infty} \phi_i(P) \phi_i(Q) . \quad (24-8)$$

¹ Provided H is separable; cf. (Meschkowski, 1962, p.48).

We could also have defined the kernel function by (24-8), supposing that the series converges for all P and $Q \in B$. It is straightforward to show by direct calculation that the kernel (24-8) possesses the reproducing property (24-2): expand $f(P)$ with respect to the basis ϕ_i :

$$f(P) = \sum_{i=1}^{\infty} f_i \phi_i(P), \quad (24-9)$$

and substitute (24-8) into (24-2):

$$\begin{aligned} (f(P), K(P, Q))_P &= (f(P), \sum \phi_i(P) \phi_i(Q))_P \\ &= \sum \phi_i(Q) (f(P), \phi_i(P))_P \\ &= \sum \phi_i(Q) f_i = f(Q), \end{aligned}$$

in view of (24-9) and (4-38).

Functionals and the dual space. In sec. 5 we have met the notion of a linear functional in a normed space H ,

$$Lf = l, \quad (24-10)$$

which associates to an element $f \in H$ a real number l , and we have defined its norm $\|L\|$ by means of (5-13).

The linear functionals L with the norm $\|L\|$ form themselves a normed space, which is called the *dual space* of H and denoted by H' .

For a Hilbert space, each bounded linear functional Lf can be represented as an inner product of f with a certain element $h \in H$

$$Lf = (h, f) \quad (24-11)$$

where, as in secs. 4 and 5, we omit the argument in the inner product, writing

$$(h, f) = (h(P), f(P))_P. \quad (24-12)$$

The element $h \in H$ corresponding in this way to a functional $L \in H'$ is called the *representer* of L .

By (4-59), the norms are equal:

$$\|L\|' = \|h\| ; \quad (24-13)$$

henceforth we shall write $\|L\|'$ for the norm of the functional L , rather than $\|L\|$, to indicate that $L \in H'$.

Therefore, (24-11) defines an "isometric isomorphism" between the spaces H' and H : to each element $L \in H'$ there corresponds an element $h \in H$ and vice versa (isomorphism), and the norms of corresponding elements are equal by (24-13) (isometry).

We may also in a natural way introduce an inner product in H' : we define it by

$$(L_1, L_2)' = (h_1, h_2) \quad (24-14)$$

as the inner product of the representers h_1 and h_2 in H . The norms are expressed in terms of inner products

$$\|L\|'^2 = (L, L)' , \quad (24-15)$$

$$\|h\|^2 = (h, h) \quad (24-16)$$

as usual, and (24-13) then follows from (24-14).

What has been said so far holds for linear functionals in any Hilbert space H . If, in addition, H has a reproducing kernel $K(P, Q)$, then the representer h of a functional L has a very simple form:

$$h(P) = L^Q K(P, Q) , \quad (24-17)$$

L^Q denoting the fact that the functional L acts on $K(P, Q)$ as a function of Q ; this notation has already been used in sec. 11. In fact,

$$\begin{aligned} (h, f) &= (f, h) = (f(P), h(P))_P = \\ &= (f(P), L^Q K(P, Q)) = L^Q (f(P), K(P, Q))_P = \\ &= L^Q f(Q) \end{aligned}$$

by (24-2), so that $Lf = (h, f)$ as was to be shown.

The norm of L is given by

$$\begin{aligned}\|L\|^2 &= (h, h) = (h(P), L^Q K(P, Q))_P = \\ &= L^Q (h(P), K(P, Q))_P = L^Q h(Q) ,\end{aligned}$$

in view of (24-17). Thus we have

$$\|L\|^2 = Lh . \quad (24-18)$$

Since

$$Lh = L^P h(P) = L^P L^Q K(P, Q) ,$$

we finally get

$$\|L\|^2 = L^P L^Q K(P, Q) . \quad (24-19)$$

Similarly we find for the inner product

$$(L_1, L_2)' = L_1^P L_2^Q K(P, Q) . \quad (24-20)$$

The evaluation functional. The evaluation functional, or delta functional, introduced by (4-57), associates to a function f its value at a particular point P :

$$\delta_P f = f(P) . \quad (24-21)$$

In a general Hilbert space, the functional δ_P is unbounded (cf. p.38); Hilbert spaces with kernel functions are characterized by the fact that *the evaluation functional is bounded*. In fact, the inner product satisfies the Schwarz inequality

$$(f, g) \leq \|f\| \|g\| , \quad (24-22)$$

which is a consequence of the properties (5-10); geometrically it means that the cosine of the angle α between the vectors f and g ,

$$\cos \alpha = \frac{(f, g)}{\|f\| \|g\|} , \quad (24-23)$$

is not greater than 1 . Applying (24-22) to (24-2) we get

$$|f(Q)|^2 \leq (f(P), f(P))_P \cdot (K(P, Q), K(P, Q))_P = \|f\|^2 \cdot K(Q, Q)$$

using (24-3), so that

$$|f(Q)| \leq C \|f\| \quad (24-24)$$

with

$$C = \sqrt{K(Q, Q)} , \quad (24-25)$$

which by (5-13) means that the evaluation functional

$$\delta_Q f = f(Q) \quad (24-26)$$

is bounded.

The representer of the evaluation functional is, by (24-17) and (24-26),

$$h(P) = \delta_Q K(P, Q) = K(P, Q) , \quad (24-27)$$

which is simply the reproducing kernel as a function of P , the "evaluation point" Q being held fixed.

Reproducing kernels in a Euclidean space. To make these abstract notions more concrete, we consider the simplest case in which the space H is not infinitely-dimensional Hilbert space but Euclidean space R^n . The elements of R^n are vectors

$$x = [x_1 \ x_2 \ \dots \ x_n]^T \quad (24-28)$$

(we regard them as column vectors, hence the transposition T), and the kernel is the symmetric, positive definite matrix K . The arguments P, Q become now indices i, j .

Take first

$$K = I = [\delta_{ij}] , \quad (24-29)$$

the unit matrix. The inner product is then simply given by

$$(x, y) = \sum_{i=1}^n x_i y_i , \quad (24-30)$$

and the reproducing property (24-2) becomes

$$f_j = \sum_{i=1}^n f_i \delta_{ij} , \quad (24-31)$$

which is (4-19). Linear functionals L are linear forms

$$Lx = \sum_{i=1}^n l_i x_i , \quad (24-32)$$

with coefficients l_i ; these coefficients form a vector which is simply the representer h of L : if

$$h = [h_1 \ h_2 \ \dots \ h_n]^T \quad (24-33)$$

then

$$h_i = l_i ; \quad (24-34)$$

this latter formula may be regarded as a consequence of (24-17):

$$h_i = \sum_{j=1}^n l_j \delta_{ij} = l_i .$$

So far, all is almost trivial. Let now K be an arbitrary positive definite regular $n \times n$ matrix:

$$K = [K_{ij}] . \quad (24-35)$$

The inner product of x and y is then defined by

$$(x, y) = x^T K^{-1} y , \quad (24-36)$$

K^{-1} being the inverse matrix of K . With the notation

$$K^{-1} = G = [g_{ij}] \quad (24-37)$$

this becomes

$$(x, y) = x^T G y. \quad (24-38)$$

With this definition of the inner product, the reproducing property holds. In fact, (24-2) becomes

$$x = (K, x) = K G x = K K^{-1} x = x. \quad (24-39)$$

Writing the linear form (24-32) as

$$Lx = l^T x, \quad (24-40)$$

we see that the representer h of L is not directly the vector $l = [l_i]$ but is related to it by

$$h = K l. \quad (24-41)$$

In fact,

$$(h, x) = (x, h) = x^T G h = x^T K^{-1} h = x^T K^{-1} K l = x^T l = l^T x,$$

in agreement with (24-40).

For the reader familiar with tensor calculus as used, for instance, in (Tienstra, 1956) or in (Hotine, 1969) we note that the elements of H are here called contravariant vectors and denoted by upper indices, for instance, x^i or y^i . The elements of the dual space H' , the linear forms, are the covariant vectors, denoted by lower indices, for instance, l_i . Then the inverse kernel (24-37) forms the covariant metric tensor g_{ij} , and (24-38) becomes for $x = y$:

$$\|x\|^2 = (x, x) = g_{ij} x^i x^j, \quad (24-42)$$

which is the fundamental metric form; summation is conventionally implied

if an index occurs once as a superscript and once as a subscript. The kernel (24-35) forms the contravariant metric tensor g^{ij} :

$$K = [g^{ij}] . \quad (24-43)$$

The linear form (24-40) becomes

$$Lx = l_i x^i , \quad (24-44)$$

and the representer $h^i = l^i$ (note the superscript!) is related to l_i by (24-41), which becomes

$$l^i = g^{ij} l_j . \quad (24-45)$$

In this way we recover all the usual manipulations of tensor algebra. The reader not familiar with this subject may disregard the relations just mentioned, except for one geometric fact. The usual Euclidean metric

$$\|x\|^2 = x^T x , \quad (24-46)$$

corresponding to $K = I$, holds for rectangular coordinate axes in R^n whose unit vectors form an orthonormal system. The metric with a general kernel matrix K , given by (24-42),

$$\|x\|^2 = x^T K^{-1} x = x^T G x , \quad (24-47)$$

corresponds to rectilinear coordinate axes whose base vectors--in R^3 , the vectors $[1,0,0]^T$, $[0,1,0]^T$, $[0,0,1]^T$ and similarly in R^n --do not constitute an orthonormal system: the coordinate axes may not form right angles with each other, or the unit vectors may not have equal length, or both. If K is a diagonal matrix, then the coordinate axes are orthogonal but the base vectors have different lengths; if K is nondiagonal, then we have oblique (nonorthogonal) axes.

How does this generalize to Hilbert space? The unit matrix corresponds to the delta function (4-22), in our notation $\delta(P,Q)$, which is not a reproducing kernel function because $\delta(P,P)$ is not finite (the "delta function" is not even an ordinary function). The use of a general kernel function corresponds, so to speak, to coordinate axes in Hilbert space whose base vectors do not form an orthonormal system.

Harmonic kernel functions. Take now H to be the space of all functions in R^3 harmonic in a region B which is the outside of a sphere of radius R_B . This sphere $r = R_B$ is assumed to have a radius slightly smaller than the mean radius of the earth $R = 6371$ km; it is the "Bjerhammar sphere" already introduced in sec. 8.

The kernel function is supposed to have the form (10-9),

$$K(P, Q) = \sum_{n=0}^{\infty} k_n \left(\frac{R^2}{rr'} \right)^{n+1} P_n(\cos \psi) \quad (24-48)$$

where P and Q have the spherical coordinates (r, θ, λ) and (r', θ', λ') and

$$\cos \psi = \cos \theta \cos \theta' + \sin \theta \sin \theta' \cos(\lambda' - \lambda) \quad (24-49)$$

as usual; for the sake of generality, we start the summation with $n = 0$. The coefficients k_n are arbitrary but so that the series converges in the region B and, most importantly, all $k_n \geq 0$.

As base functions ϕ_i we take the functions

$$\phi_i(P) = \begin{cases} \sqrt{\frac{k_n}{2n+1}} \left(\frac{R}{r} \right)^{n+1} \bar{R}_{nm}(\theta, \lambda) \\ \sqrt{\frac{k_n}{2n+1}} \left(\frac{R}{r} \right)^{n+1} \bar{S}_{nm}(\theta, \lambda) \end{cases} \quad (24-50)$$

arranged in some linear order, e.g., as in (21-2) but starting with $n = 0$; the functions \bar{R}_{nm} and \bar{S}_{nm} are the fully normalized Legendre harmonics (3-26). In agreement with (24-8) we form

$$\sum_{i=1}^{\infty} \phi_i(P) \phi_i(Q) = \sum_{n=0}^{\infty} \sum_{m=0}^n \frac{k_n}{2n+1} \left(\frac{R^2}{rr'} \right)^{n+1} (\bar{R}_{nm} \bar{R}'_{nm} + \bar{S}_{nm} \bar{S}'_{nm}), \quad (24-51)$$

where $\bar{R}_{nm} = \bar{R}_{nm}(\theta, \lambda)$ and $\bar{R}'_{nm} = \bar{R}_{nm}(\theta', \lambda')$. In view of (3-30), this expression is equal to the kernel (24-48) as it should be.

The inner product of two functions f and g in H is defined as follows. Expand the function $f(\theta, \lambda) = f(P)$ on the sphere $r = R$ into the series (3-28):

$$f = \sum_{n=0}^{\infty} \sum_{m=0}^n (\bar{A}_{nm} \bar{R}_{nm} + \bar{B}_{nm} \bar{S}_{nm}), \quad (24-52)$$

and similarly for g :

$$g = \sum_{n=0}^{\infty} \sum_{m=0}^n (\bar{C}_{nm} \bar{R}_{nm} + \bar{D}_{nm} \bar{S}_{nm}) . \quad (24-53)$$

Then we define

$$(f, g) = \sum_{n=0}^{\infty} \frac{2n+1}{k_n} \sum_{m=0}^n (\bar{A}_{nm} \bar{C}_{nm} + \bar{B}_{nm} \bar{D}_{nm}) . \quad (24-54)$$

The norm is then

$$\|f\|^2 = \sum_{n=0}^{\infty} \frac{2n+1}{k_n} \sum_{m=0}^n (\bar{A}_{nm}^2 + \bar{B}_{nm}^2) . \quad (24-55)$$

The verification that the reproducing property (24-2) is satisfied for this inner product definition is straightforward: putting

$$g(P) = K(P, Q) \quad (Q \text{ fixed})$$

we have for its coefficients corresponding to (24-53), by (24-51),

$$\bar{C}_{nm} = \frac{k_n}{2n+1} \bar{R}'_{nm} , \quad \bar{D}_{nm} = \frac{k_n}{2n+1} \bar{S}'_{nm} ,$$

and hence by (24-54)

$$\begin{aligned} (f(P), K(P, Q))_P &= (f, g) = \\ &= \sum_{n=0}^{\infty} \frac{2n+1}{k_n} \frac{k_n}{2n+1} \sum_{m=0}^n (\bar{A}_{nm} \bar{R}'_{nm} + \bar{B}_{nm} \bar{S}'_{nm}) \\ &= f(Q) . \end{aligned}$$

It should be pointed out that, in the present case, the inner product (f, g) is not simply defined by an integral as it was in the case of L_2 , eq. (4-28).

We note that the positivity of the "spectrum" formed by the spherical-harmonic coefficients k_n is essential for $K(P, Q)$ to be an admissible kernel function: only then is the representation (24-51) by real functions (24-50) possible. On the other hand, kernel functions are positive definite

according to (24-5). Thus we see that the condition (24-5) and the positivity of the spectrum are closely related.

25. COLLOCATION AND HILBERT SPACE

The formulas of least-squares collocation as introduced in Part B of the book find a natural geometrical interpretation in Hilbert space with a kernel function. In this section we shall consider the Hilbert space interpretation of "pure" collocation without noise and systematic parameters, as discussed in secs. 11 and 12.

The minimum norm property. The q given linear functionals (11-3),

$$L_i T = l_i, \quad i = 1, 2, \dots, q, \quad (25-1)$$

define a hyperplane D of H of codimension q .

What does this mean? Consider the case of R^n . Then a linear functional has the form (4-52). This is the equation of a hyperplane of dimension $n-1$, for which we may also say, of *codimension* 1. As long as $q < n$, q linear functionals define a subspace of dimension $n-q$, which is the intersection of q hyperplanes corresponding to the given functionals. Instead of dimension $n-q$, we may also speak of codimension q , and this expression has the advantage that it can also be used in Hilbert space where n is infinite.

In (25-1), the potential T is unknown but the numerical values l_i of the q functionals $L_i T$ are given. All possible functions T compatible with the observations l_i must satisfy the system (25-1). In geometric terms, considering T as a vector in Hilbert space, all possible solutions of (25-1) must lie in the hyperplane D . Cf. Fig. 25.1, which corresponds to the simplified situation that H is a plane. All possible solutions \bar{T} lie in D ; among them there is one, \hat{T} , for which the norm (geometrically, the "length") is minimum:

$$\|\bar{T}\| \geq \|\hat{T}\|. \quad (25-2)$$

This solution \hat{T} is orthogonal to the hyperplane D , and it is precisely the solution (11-6) given by least-squares collocation. This can be seen as follows.

The solution \hat{T} is a linear combination of base functions (12-12),

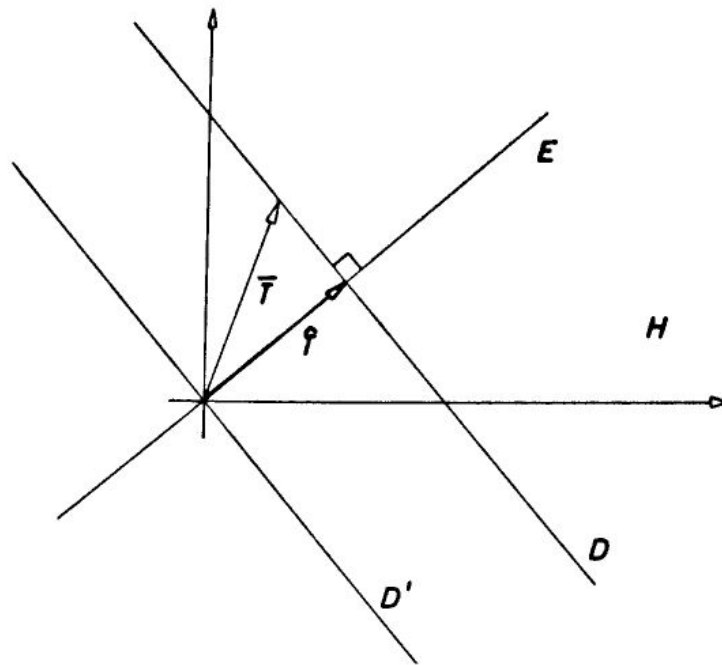


FIGURE 25.1. Minimum norm as minimum distance.

$$\phi_i(P) = L_i^Q K(P, Q) , \quad (25-3)$$

so that

$$\bar{T}(P) = \sum_{i=1}^q b_i \phi_i(P) . \quad (25-4)$$

These q base functions span a subspace E of dimension q , and $\bar{T}(P)$, as a linear combination of base functions, belongs to this subspace.

Let us now show that E is orthogonal to the hyperplane D defined by (25-1), or, which is the same, to the hyperplane D' through the origin and parallel to D (Fig. 25.1); D' is a subspace of H . The subspace D' is thus defined by the formulas

$$L_i \bar{T} = 0 , \quad i = 1, 2, \dots, q . \quad (25-5)$$

Let s_0 be an arbitrary element of D' , so that

$$L_i s_0 = 0 . \quad (25-6)$$

Then $D' \perp E$ if and only if

$$(s_0, \phi_i) = 0 , \quad i = 1, 2, \dots, q , \quad (25-7)$$

since E is spanned by the elements ϕ_i . Now

$$\begin{aligned} (s_0, \phi_i) &= (s_0(P), L_i^Q K(P, Q))_P \\ &= L_i^Q (s_0(P), K(P, Q))_P \\ &= L_i^Q s_0(Q) = L_i s_0 = 0 ; \end{aligned}$$

here we have used (25-3), (24-2), and (25-6). This proves (25-7), so that the subspaces D' and E are orthogonal; together they span the whole space H .

Thus the estimate \hat{T} will be orthogonal to D and so have shortest length $\|\hat{T}\|$, provided it has the form (25-4) of an element of E . This can also be shown directly. Write any estimate \bar{T} satisfying (25-1) in the form

$$\bar{T} = \hat{T} + T_0 , \quad (25-8)$$

where $\hat{T} \in E$ and $T_0 \in D'$. Then

$$\begin{aligned} (\bar{T}, \bar{T}) &= (\hat{T} + T_0, \hat{T} + T_0) \\ &= (\hat{T}, \hat{T}) + 2(\hat{T}, T_0) + (T_0, T_0) . \end{aligned}$$

Now $(\hat{T}, T_0) = 0$ because $D' \perp E$, so that

$$\|\bar{T}\|^2 = \|\hat{T}\|^2 + \|T_0\|^2 \geq \|\hat{T}\|^2 , \quad (25-9)$$

which proves the minimum norm property (25-2) of the least-squares collocation estimate.

An explicit solution. The coefficients b_i are determined by substituting (25-4) into (25-1). In geometrical terms, the least-squares estimate \hat{T} is the (unique) intersection point of D and E (Fig.25.1).

Algebraically we may proceed as follows. Using the abbreviation (11-5),

$$B = [L_i] , \quad (25-10)$$

we may write (25-1) in the form (11-4)

$$BT = 1 \quad (25-11)$$

where 1 is a q -vector and B is a linear operator comprising the functionals L_i according to (11-5); in the terminology of sec. 5, cf. (5-28), we may say that B is a mapping

$$B : H \rightarrow R^q \quad (25-12)$$

of a function T into a vector 1 consisting of q real numbers $1_i = L_i T$.

Using (25-10), we may write (11-13),

$$C_{Pi} = L_i^Q K(P, Q) = \phi_i(P) , \quad (25-13)$$

in the form

$$[\phi_i] = [C_{Pi}] = BK , \quad (25-14)$$

so that (25-4) becomes

$$\hat{T} = (BK)^T b . \quad (25-15)$$

The insertion into (25-11) gives

$$B(BK)^T b = 1 ,$$

which determines the coefficients b :

$$b = [B(BK)^T]^{-1} 1 , \quad (25-16)$$

so that (25-15) becomes finally

$$\hat{T} = (BK)^T [B(BK)^T]^{-1} \quad (25-17)$$

This equation is precisely (11-6), in view of (25-14) and (11-14) which can be written

$$[C_{ij}] = B(BK)^T \quad (25-18)$$

With the abbreviations

$$C_{s1} = (BK)^T \quad C_{11} = B(BK)^T, \quad (25-19)$$

this takes the usual Wiener-Kolmogorov form

$$\hat{T} = C_{s1} C_{11}^{-1} \quad (25-20)$$

Of the wealth of geometrical and functional relationships of least-squares collocation we mention still one. Using (25-11), we may write (25-17) as

$$\hat{T} = P_E T, \quad (25-21)$$

where

$$P_E = (BK)^T [B(BK)^T]^{-1} B \quad (25-22)$$

is the projection operator onto the subspace E :

$$P_E : H \rightarrow E. \quad (25-23)$$

It projects any element $s \in H$ orthogonally onto E by $\hat{s} = P_E s$ (Fig.25.2).

Relation to sec. 21. Suppose that the elements of the Hilbert space H are not functions but infinite vectors

$$s = [s_1 \ s_2 \ s_3 \ \dots]^T, \quad (25-24)$$

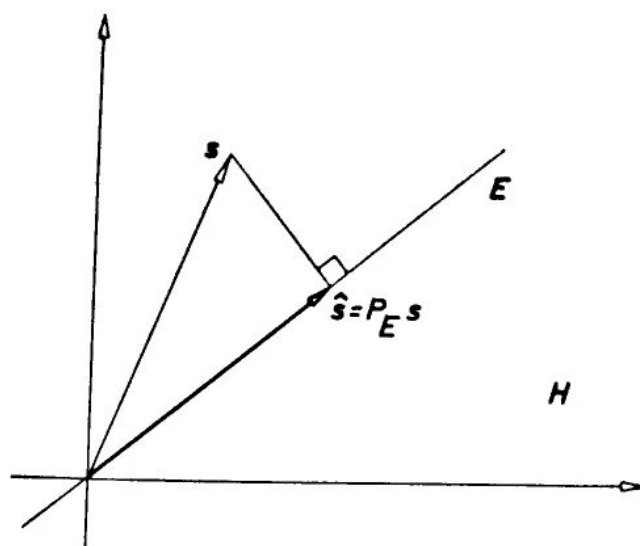


FIGURE 25.2. The projection operator P_E .

cf. (4-6). Then the kernel K is a symmetric infinite matrix. This is completely analogous to the finite-dimensional case; cf. (24-28) and (24-35).

The operator B is then a $q \times \infty$ matrix, and

$$(BK)^T = KB^T, \quad (25-25)$$

so that (25-17) becomes

$$\hat{s} = KB^T(BKB^T)^{-1}l. \quad (25-26)$$

This is precisely the case discussed in sec. 21. The spherical-harmonic coefficients (21-2) form an infinite vector (25-24), and the least-squares collocation formula (21-33) for errorless observations is (25-26).

In sec. 4 we have mentioned the isomorphism between function space L_2 and sequence space l_2 , defined by an expansion (4-35) into a series of base functions. In the present case of a Hilbert space H with a kernel function matters are completely analogous. We may expand a harmonic function $f(P)$ in space into a series of base functions (24-50):

$$f(P) = \sum_{i=1}^{\infty} f_i \phi_i(P) . \quad (25-27)$$

On the sphere $r = R$, this expansion reduces to (24-52), and the comparison with (24-50) shows that

$$f_i = \begin{cases} \frac{\sqrt{2n+1}}{k_n} \bar{A}_{nm} , \\ \frac{\sqrt{2n+1}}{k_n} \bar{B}_{nm} , \end{cases} \quad (25-28)$$

arranged in the same linear order as the functions (24-50).

For $f = T$ this becomes

$$T(P) = \sum_{i=1}^{\infty} s_i \phi_i(P) , \quad (25-29)$$

so that the sequence (25-24) is the equivalent of T in a Hilbert space l_2 of infinite vectors (25-24). Eq. (25-29) thus defines an isomorphism between H and l_2 :

$$H \rightarrow l_2 : T \rightarrow s \quad (25-30)$$

which can even be shown to be isometric:

$$\|T\|^2 = \|s\|^2 = (s, s) = s^T s . \quad (25-31)$$

In fact, (24-55) becomes with (25-28)

$$\|f\|^2 = \sum_{i=1}^{\infty} f_i^2 = f^T f , \quad (25-32)$$

which is the l_2 norm (4-7). Thus, the matrix K reduces to the infinite unit matrix I .

In sec. 21 we have expanded, not with respect to the base functions (24-50), but with respect to the fully normalized harmonics

$$\left(\frac{R}{r}\right)^{n+1} R_{nm}(\theta, \lambda), \quad \left(\frac{R}{r}\right)^{n+1} S_{nm}(\theta, \lambda), \quad (25-33)$$

which differ from (24-50) by the factor $\sqrt{k_n}/(2n+1)$. Therefore, the kernel matrix K has not been the unit matrix but a diagonal matrix (21-23) with elements $k_n/(2n+1)$, apart from a factor; cf. (21-24). The infinite vector s has consisted of the coefficients of the base functions (25-33), and hence the norm $\|s\|$ is given by

$$\|s\|^2 = s^T K^{-1} s, \quad (25-34)$$

which is clearly the same as (24-55), since K^{-1} is a diagonal matrix with elements $(2n+1)/k_n$.

Still there is isometry:

$$\|T\| = \|s\| = (s^T K^{-1} s)^{1/2}, \quad (25-35)$$

and hence in (21-37) we recognize our familiar minimum norm condition (25-2).

The minimum norm. Let us now explicitly evaluate the norm $\|\hat{T}\|$ of the least-squares solution (25-17). Using (25-3), (25-4) and (24-2) we have

$$\begin{aligned} \|\hat{T}\|^2 &= (\hat{T}, \hat{T}) = (\hat{T}(P), \sum_{i=1}^q b_i L_i^Q K(P, Q))_P \\ &= \sum_{i=1}^q b_i L_i^Q (\hat{T}(P), K(P, Q))_P \\ &= \sum_{i=1}^q b_i L_i^Q \hat{T}(Q) = \sum_{i=1}^q b_i L_i^Q \sum_{j=1}^q b_j L_j^R K(Q, R) \\ &= \sum_{ij} b_i b_j L_i^Q L_j^R K(Q, R). \end{aligned}$$

By means of (11-14) this can be written as

$$\|\hat{T}\|^2 = b^T C_{11} b, \quad (25-36)$$

which by (25-16) and (25-19) finally becomes

$$\|\hat{T}\|^2 = 1^T C_{11}^{-1} 1. \quad (25-37)$$

This simple formula expresses the norm in terms of the data. It looks similar to (25-35), but note that \mathbf{l} is a q -vector, whereas \mathbf{s} is an infinite vector.

Covariances as inner products of functionals. Let us now turn to the other property of the least-squares estimate, namely minimum error variance. For this purpose we first consider the inner product of two linear functionals, L_1 and L_2 , which are elements of the dual space H' (sec.24). By (24-20) we have

$$(L_1, L_2)' = L_1^P L_2^Q K(P, Q) . \quad (25-38)$$

On the other hand, eqs. (11-11), (11-12), and (11-14) show that, if

$$L_1^T = \mathbf{l}_1 , \quad L_2^T = \mathbf{l}_2 , \quad (25-39)$$

then

$$\text{cov}(\mathbf{l}_1, \mathbf{l}_2) = M\{\mathbf{l}_1 \mathbf{l}_2\} = L_1^P L_2^Q K(P, Q) . \quad (25-40)$$

We thus see that

$$(L_1, L_2)' = \text{cov}(\mathbf{l}_1, \mathbf{l}_2) = M\{\mathbf{l}_1 \mathbf{l}_2\} . \quad (25-41)$$

This fundamental result may be expressed as follows. *If we identify the kernel function $K(P, Q)$ with the covariance function, then the covariances of the values of two linear functionals of T are the inner products of the two functionals in the dual space H' .*

For $L_2 = L_1$ we get

$$\|L_1\|^2 = \text{var}(\mathbf{l}_1) = M\{\mathbf{l}_1^2\} , \quad (25-42)$$

the variance of a functional is the square of the norm in H' .

In this way, statistics is related to Hilbert space geometry: we get a geometrical interpretation of variances and covariances.

The geometry of minimum error variance. Consider a signal

$$\mathbf{s}_k = \mathbf{S}_k^T , \quad (25-43)$$

which is a linear functional S_k of T , and let m of these signals form the vector s ; cf. (11-25) and (11-26). The estimate of s by least-squares collocation is given by (11-23) or (11-27):

$$\hat{s} = C_{s1} C_{11}^{-1} l . \quad (25-44)$$

We put

$$h = C_{s1} C_{11}^{-1} \quad (25-45)$$

where

$$h = [h_{ki}] \quad (25-46)$$

is a $m \times q$ matrix, and write (25-44) in the form

$$\hat{s} = h l \quad (25-47)$$

or

$$\hat{s}_k = \sum_{i=1}^q h_{ki} l_i , \quad (25-48)$$

expressing estimates of the unknown quantities s_k as linear combinations of the given quantities l_i .

In sec. 9 we have asked the question: how is the matrix h to be selected in order to minimize the standard errors of all estimated signals \hat{s}_k . The answer was that it must be the least-squares estimate (25-45).

What is the geometrical interpretation of these standard errors of estimation? The individual error of \hat{s}_k is, by (9-14),

$$\epsilon_k = \hat{s}_k - s_k = (\hat{S}_k - S_k)T . \quad (25-49)$$

Hence the error variance (11-35) becomes, by (25-42),

$$\sigma_k^2 = \text{var}(\epsilon_k) = M\{\epsilon_k^2\} = M\{(\hat{s}_k - s_k)^2\} = \|\hat{S}_k - S_k\|^2 . \quad (25-50)$$

Thus the standard error of estimation

$$\sigma_k = \|\hat{S}_k - S_k\|' \quad (25-51)$$

is nothing else than the norm of the "error functional" $\hat{S}_k - S_k$ in H' .
The minimization of this error norm,

$$\|\hat{S}_k - S_k\|' = \text{minimum} , \quad (25-52)$$

leads again to the least-squares estimate (25-45), in exactly the same way as in sec. 9, via expressions such as (9-19): in fact, the covariances entering in these expressions may now be interpreted as inner products in H' .

The geometrical interpretation of this algebraic procedure is as follows (Krarup, 1978, pp.200-201). We assume $k = 1$ which is no restriction since all \hat{S}_k are obtained independently. We write $S_1 = S$ and $h_{11} = h_1$; thus (25-48) gives

$$\hat{S} = \sum_{i=1}^q h_i L_i . \quad (25-53)$$

We have to determine the coefficients h_i in such a way as to minimize the error norm (25-52),

$$\|\hat{S} - S\|' = \text{minimum} . \quad (25-54)$$

The problem is now formulated completely in terms of the geometry of the dual space H' (Fig.25.3).

The given elements $L_i \in H'$ form a linear subspace of H' of dimension q ; we call it H'_L . The condition (25-54) now corresponds to the orthogonal projection of S onto the subspace H'_L . In other terms, the functional $\hat{S} - S$ is orthogonal to all functionals L_j spanning H'_L :

$$(\hat{S} - S, L_j)' = 0 \quad (25-55)$$

or

$$(\hat{S}, L_j)' = (S, L_j)' .$$

If we have again m functionals S_k , then this equation must be satisfied for all of them:

$$(\hat{S}_k, L_j)' = (S_k, L_j)' . \quad (25-56)$$

By (25-48),

$$\hat{S}_k = \sum_{i=1}^q h_{ki} L_i , \quad (25-57)$$

this becomes

$$\sum_{i=1}^q h_{ki} (L_i, L_j)' = (S_k, L_j)' . \quad (25-58)$$

In matrix notation and using the relation between inner products and covariances, this is

$$hC_{11} = C_{s1} , \quad (25-59)$$

from which (25-45) follows.

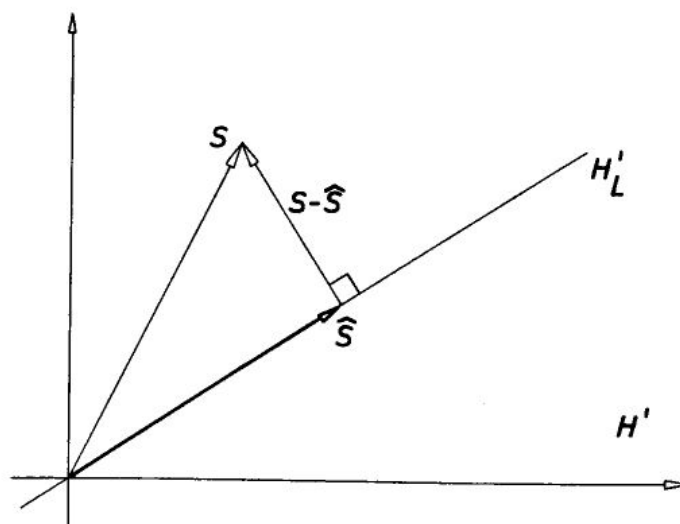


FIGURE 25.3. *Geometry of minimum error norm.*

The norm of T. An interesting consequence of the use of the covariance function as the kernel function has been pointed out by Tscherning (1977). Consider the norm of the anomalous potential T . Using the spherical-harmonic expansion (10-6) and the expression (24-55) of the norm, we have

$$\|T\|^2 = \sum_{n=2}^{\infty} \frac{2n+1}{k_n} \sum_{m=0}^n \left(\bar{a}_{nm}^2 + \bar{b}_{nm}^2 \right). \quad (25-60)$$

If we choose $K(P,Q)$ as the covariance function, then k_n is given by (10-8) and this expression becomes

$$\|T\|^2 = \sum_{n=2}^{\infty} (2n+1).$$

This result means that the norm of T is not finite, so that the potential T itself does not belong to the Hilbert space H .

This fact, which least-squares collocation shares with the prediction theory of stochastic processes, appears to be a mathematical subtlety rather than a practical difficulty. It is true that when $\|T\|$ is not finite, the simple convergence proof given in (Moritz, 1976a, p.14) cannot be applied: there it has been shown that, if $\|T\|$ is finite, then the solution for least-squares interpolation tends to the true function as the density of data points increases indefinitely. Practically, however, the covariance function can never be determined precisely, because this would require the knowledge of T or Δg over the whole terrestrial sphere--in this ideal case we should know the gravity field without needing collocation!--, and empirical covariance functions can always be modified without harm in such a way that (25-60) becomes finite. For instance, if

$$k_n = \sum_{m=0}^n (\bar{a}_{nm}^2 + \bar{b}_{nm}^2)$$

were exactly known, it would suffice to change k_n to $k_n(1+\epsilon)^n$, ϵ being positive and as small as we like, to change the divergence of (25-60) into convergence.

Concluding remarks. We have seen that both classical minimum principles due to Gauss, least squares and minimum variance, can for least-squares collocation be geometrically interpreted in Hilbert space. The first principle becomes a minimum norm condition (25-2); in geometrical terms we seek the distance of the "observation hyperplane" D from the origin (Fig.25.1).

This interpretation holds whether we identify the kernel function with the covariance function or not.

The second Gaussian principle, minimum variance, leads to a "best" estimate in a statistical sense which, by identifying the kernel function with the covariance function, can be geometrically interpreted as a minimum error norm approximation in H' according to (25-54). This geometrical property holds even when we work with an arbitrary kernel function ("analytical collocation", cf. sec.12). In this case we do not, however, have a statistical interpretation in terms of minimum variance.

By (24-55), the norm will be small if the spherical-harmonic coefficients are small, that is, if the function is smooth. In this sense, minimum norm means greatest smoothness, and we may say that least-squares collocation gives the smoothest gravitational field that is compatible with the given data. This is certainly a physically reasonable property because it avoids spurious irregularities without empirical basis, in line with the maxim (expressed by Sir Harold Jeffreys): "When in doubt, smooth".

An analogous geometric interpretation of stochastic processes in terms of a Hilbert space with the covariance function as the kernel function has been given by Parzen (1961). The geometric treatment of least-squares collocation is due to Krarup (1969) and has been used by Tscherning (1975b) and others. The corresponding geometric situation for least-squares adjustment has been illustrated by means of a simple example in (Moritz, 1966). See also (Meissl, 1976) and (Dermanis, 1977).

Equation (25-11) has the form of a system of condition equations for least-squares adjustment in Hilbert space. This is particularly evident for a Hilbert space of sequences, where (25-11) is given by (21-32). Physically we have the difference that here our statistical variables, the signals, are quantities of the anomalous gravitational field, whereas in adjustment the random variables are observational errors. Mathematically the structure is the same, except that in adjustment the space is finite-dimensional, whereas in prediction and collocation we have infinitely-dimensional Hilbert space. The interpretation of prediction as an adjustment in Hilbert space was pointed out by Krarup (1969, pp.39-41).

In the present section we have limited ourselves to the case of "pure collocation", without observational errors and without parameters. The general case of least-squares collocation will be taken up in secs. 29 and 30.

26. GEODETIC MEASUREMENTS AND THEIR REPRESENTATION

There are essentially two possible approaches to physical geodesy (as also to other natural sciences): they might be called the model approach and the operational approach. Essentially, the first approach starts from a theory, the second from the observations. Obviously, the two approaches are closely related to the deductive method and the inductive method in the natural sciences.

In the *model approach*, one starts from a mathematical model or from a theory and then tries to fit this model to reality, for instance by determining the parameters of this model from observations. The classical geodetic example are the centuries-old attempts to determine the parameters of an earth ellipsoid by observation, from the old grade measurements to modern satellite observations.

Perhaps the most elaborate form of this model approach is the boundary-value problem of physical geodesy in the formulation of Molodensky, to be discussed later in Part D. It has a mathematically interesting and deep theory and is practically highly significant. Still, this approach also has its weaknesses: the required continuous gravity coverage is practically not realizable; on the other hand, many other important data cannot be incorporated into this theory. The model selects its data.

At present we have a great number of geodetic measurements of very different types, from terrestrial angle and distance measurements to satellite data of various kinds. The question arises: how can we use and combine all these data in the best possible way. This is the *operational approach*.

Let us summarize. In the model approach one asks: how can I best determine my model by suitable observations? In the operational approach one asks: how can I make best use of all my observations?

As a matter of fact, the two approaches do not compete with each other; each one incorporates important aspects, and the two approaches mutually complement each other.

The operational approach to physical geodesy has come up at a relatively recent date, when a huge number of measurements of new types was available and when it turned out that the classical, especially the gravimetric, approach failed to give a complete answer in view of the lack of gravity data.

In geometrical geodesy already least-squares adjustment is in the spirit of an operational approach. In physical geodesy we have least-squares collocation and similar methods ("integrated geodesy", "operational geodesy").

In the present section and the following ones we shall attempt a more rigorous and general treatment of the operational approach than we did in Part B. This treatment follows (Moritz, 1978d); we shall start from the mathematical representation of geodetic measurements.

Measurements as nonlinear functionals. Every geodetic measurement depends:

1. on one or several points in space;
2. on the earth's gravitational field.

Symbolically we may write:

$$l = F(X, V) . \quad (26-1)$$

Here l denotes the measurement under consideration, V denotes the gravitational potential, and the vector X comprises the coordinates of the points to which the measurement refers, and possibly other parameters. For instance, if we have two points P and Q and if we use rectangular coordinates xyz referred to some cartesian reference system then

$$X = [x_P \ y_P \ z_P \ x_Q \ y_Q \ z_Q]^T , \quad (26-2)$$

T denoting the transpose (as usual, vectors will be column vectors unless the contrary is stated).

The symbol F denotes any functional dependence on X and V . With respect to X , it is an ordinary function; but it is not necessarily an ordinary function of V but may involve first and higher derivatives of V , integrals, etc. In the terminology of functional analysis, F is a (nonlinear) *functional* of X and V ; cf. sec. 5.

Denote the number of components of X by p ; then X may be said to belong to p -dimensional Euclidean space R^p . The function V may be considered to belong to some set, or space, H of harmonic functions. Then, in the terminology mentioned at the end of sec. 5, the functional F is a mapping of the product space¹ $R^p \times H$ into R , the real number line:

$$F : R^p \times H \rightarrow R , \quad (26-3)$$

¹ A *product space* is defined as follows. Let U and V be two spaces, and consider two elements $u \in U$ and $v \in V$. Then the ordered pair (u, v) is, by definition, an element of the product space $U \times V$. For instance, the plane R^2 is the product space $R \times R$ where R denotes the real line, since $(x, y) \in R^2$ if $x \in R$, $y \in R$. Therefore, a product space is also called a *cartesian product*.

which, in plain language, means simply that F associates, to each harmonic function from the set H and to each vector $\in R^P$, a real number which represents the numerical value of the observation l .

More intuitively we may say that F is nothing else but a prescription for computing a number l from a given vector X and a given function V : if X and V are supposed to be known, then it must be possible to find, in an unambiguous way, the value of l . In other terms, F denotes an operation to be performed on X and V , the result of which is a real number.

Our functionals (26-1) will, in general, be nonlinear; in the next section, we shall describe how they can be linearized. The physical meaning of such functionals F will be clear from the examples given below.

Instead of the gravitational potential V , we may also use the gravity potential W , defined in the usual way by

$$W = V + \frac{1}{2} \omega^2 (x^2 + y^2), \quad (26-4)$$

ω denoting the angular velocity of the earth's rotation and the z -axis coinciding with the earth's mean axis of rotation (sec.1).

Then, instead of (26-1), we have

$$l = F(X, W). \quad (26-5)$$

As W is expressed in terms of V and X by (26-4), the relations (26-1) and (26-5) are equivalent; as a matter of fact, the letter F denotes different functionals in each of the two cases.

Let us now illustrate these abstract considerations by means of concrete examples.

Astronomical and gravimetric observations. The gravity vector \underline{g} is expressed in terms of gravity g and astronomical latitude ϕ and longitude Λ by means of the relation following from (1-14) and (1-15):

$$-\underline{g} = \begin{bmatrix} g \cos \phi \cos \Lambda \\ g \cos \phi \sin \Lambda \\ g \sin \phi \end{bmatrix} \quad (26-6)$$

(cf. Heiskanen and Moritz, 1967, p.57). On the other hand, \underline{g} is the gradient of the gravity potential W :

$$\underline{g} = \text{grad } W = \begin{bmatrix} W_x \\ W_y \\ W_z \end{bmatrix}, \quad (26-7)$$

W_x denoting the partial derivative

$$W_x = \frac{\partial W}{\partial x} \quad (26-8)$$

and similarly for W_y and W_z . Comparing (26-6) and (26-7) and solving for ϕ , Λ , and g we obtain

$$\phi = \tan^{-1} \frac{-W_z}{\sqrt{W_x^2 + W_y^2}}, \quad (26-9)$$

$$\Lambda = \tan^{-1} \frac{W_y}{W_x}, \quad (26-10)$$

$$g = \sqrt{W_x^2 + W_y^2 + W_z^2}. \quad (26-11)$$

These equations have the form (26-5): they express the observables ϕ , Λ , g in terms of the potential W , not as ordinary functions of W , but as nonlinear functionals involving the operation of differentiation. Let X denote the coordinate vector of the observation station:

$$X = \begin{bmatrix} x \\ y \\ z \end{bmatrix}. \quad (26-12)$$

Then, as W_x , W_y , W_z are functions of x, y, z , the expressions (26-9) to (26-11) do, in fact, also depend on X , in agreement with (26-5).

Angle and distance measurements. The observables: azimuth α , zenith distance ζ , and distance s between two points P and Q , can be expressed in terms of the coordinate differences

$$\begin{aligned} \Delta x &= x_Q - x_P, \\ \Delta y &= y_Q - y_P, \\ \Delta z &= z_Q - z_P, \end{aligned} \quad (26-13)$$

as follows (Heiskanen and Moritz, 1967, p.219):

$$\alpha = \tan^{-1} \frac{-\Delta x \sin \Lambda + \Delta y \cos \Lambda}{-\Delta x \sin \phi \cos \Lambda - \Delta y \sin \phi \sin \Lambda + \Delta z \cos \phi}, \quad (26-14)$$

$$\zeta = \cos^{-1} \frac{\Delta x \cos \phi \cos \Lambda + \Delta y \cos \phi \sin \Lambda + \Delta z \sin \phi}{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}, \quad (26-15)$$

$$s = \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}. \quad (26-16)$$

Again, these equations have the form (26-5); the vector X is now

$$X = [x_P \ y_P \ z_P \ x_Q \ y_Q \ z_Q]^T, \quad (26-17)$$

comprising the coordinates of both points P and Q , and the dependence on the potential W is implicit through ϕ and Λ as expressed by (26-9) and (26-10); hence α and ζ are, in fact, nonlinear functionals of W . Note that these observables depend on the target point Q only because its coordinates enter into $\Delta x, \Delta y, \Delta z$; on the observation station P they depend in the same way, but there is an additional dependence on P because ϕ and Λ , and hence W_x, W_y, W_z , refer to this point.

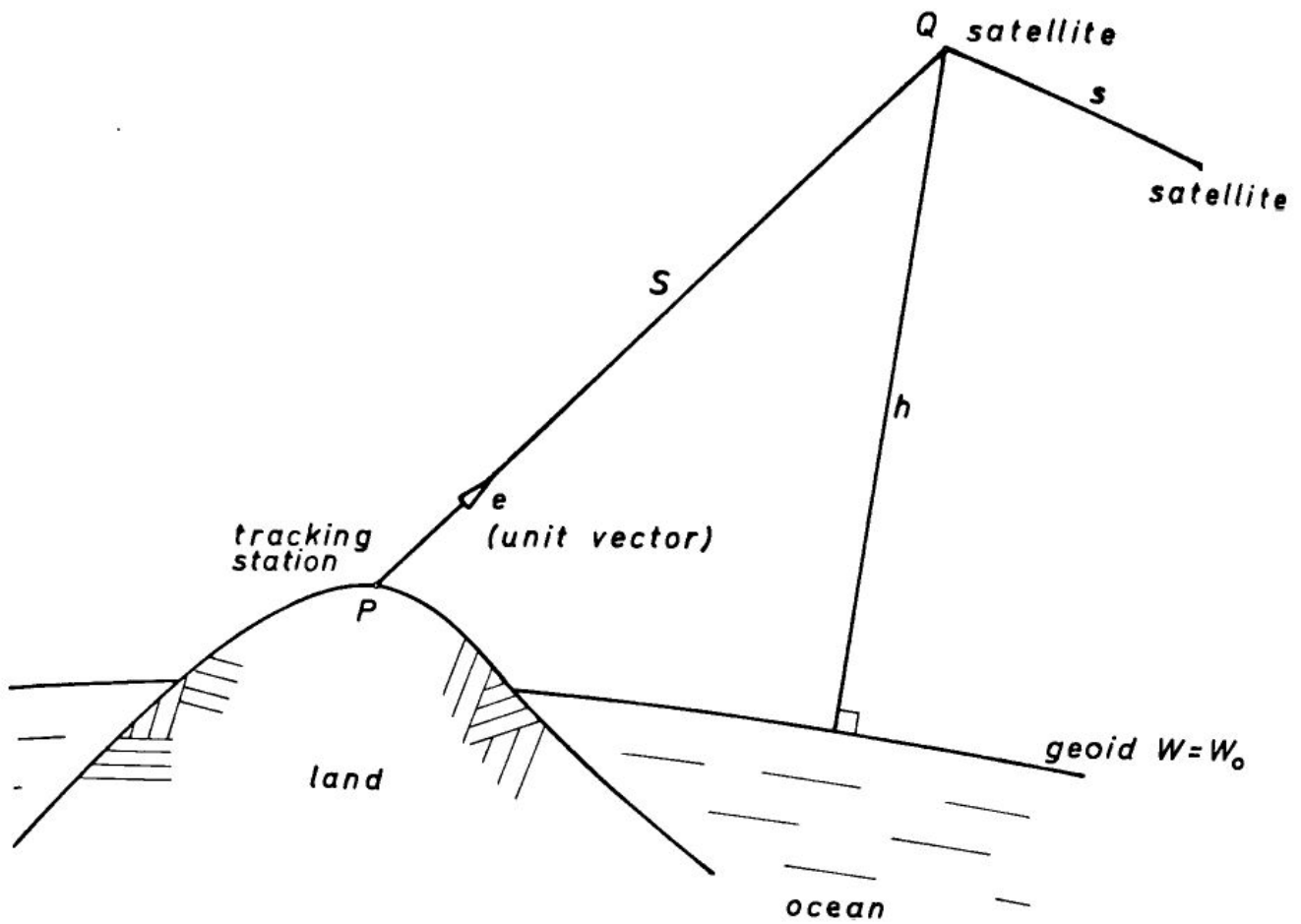
A measured horizontal angle ω may be considered as the difference between two azimuths:

$$\omega = \alpha_2 - \alpha_1, \quad (26-18)$$

measured at an observation station P to two targets Q_1 and Q_2 . Both azimuths α_1 and α_2 may be expressed by (26-14); the resulting expression for ω clearly involves the coordinates of P, Q_1 , and Q_2 , so that, in the present case, the vector X consists of 9 components, which are the coordinates of these three points; we again get a nonlinear functional of form (26-5).

It goes without saying that the functional expression for the spatial distance (26-16) is also a special case of (26-5), in which there simply is no factual dependence on W : measured straight distances between two fixed points do not depend on the gravity field.

Satellite observations. Consider a distance S measured from a ground station P to a satellite Q by laser or radar. (Cf. Fig. 26.1; for a non-technical and compact review of various techniques see (Cordova, 1977).)



e	direction observation
S	distance measurement
dS/dt	doppler observation
h	satellite altimetry
ds/dt	satellite-to-satellite tracking (doppler)

FIGURE 26.1. Satellite techniques.

Such a distance can again be represented by (26-16) but, if we operate in the orbital mode, the coordinates of Q can be further expressed by the six orbital elements p_1, p_2, \dots, p_6 of some reference orbit and the coefficients J_{nm} and K_{nm} of the expansion of the earth's gravitational potential V in terms of spherical harmonics. Thus S will have the form of some function

$$S = S(x_P, y_P, z_P; p_1, p_2, \dots, p_6; J_{nm}, K_{nm}) \quad (26-19)$$

This is a functional of form (26-1). The vector X is given now by

$$X = [x_P \ y_P \ z_P \ p_1 \ p_2 \ \dots \ p_6]^T ; \quad (26-20)$$

it comprises station coordinates and orbital parameters. The spherical-harmonic coefficients J_{nm} and K_{nm} may be expressed in terms of V by well-known integral formulas (of type of eq. (3-21)), which explains the functional dependence on V .

The change of distance S with respect to time t , that is, the *range-rate* dS/dt , can be measured by doppler observations. By integrating dS/dt with respect to t from t_1 to t_2 , one obtains distance differences $S_2 - S_1$. By photographing the satellite against the background of stars one finds the right ascension and the declination of the spatial direction PQ , or in other terms, the unit vector e of this direction (Fig. 26.1). All these observables have the same mathematical structure as (26-19): they are again functionals of type (26-1), the vector X being given by (26-20).

Satellite altimetry can be considered to measure the height h of a satellite above the geoid: the ocean surface reflects a radar signal emitted by the satellite, and under idealized conditions, this surface coincides with the geoid. We claim that h can again be expressed as a functional of type (26-1), with

$$X = [p_1 \ p_2 \ \dots \ p_6]^T . \quad (26-21)$$

This is true if it is possible, given X and the potential function $V(x,y,z)$, to compute h . In fact, assume the gravitational potential $V(x,y,z)$ to be known as a function of position at all points outside and on the earth's surface. Then the gravity potential function W is also known by (26-4), and consequently the geoid is an equipotential surface

$$W(x,y,z) = W_0 = \text{const.} \quad (26-22)$$

Now, the satellite orbit can be computed from the parameters p_k of the reference orbit and the gravitational potential V , and the position Q of the satellite along the orbit is uniquely determined by giving the corresponding instant t (which we assume to be known). Thus both the geoid and the satellite position Q are determined, and so is h , as the length of the perpendicular from Q to the geoid. Therefore, by the definition

of the functional (26-1) given above, the satellite altimeter measurement h is, in fact, such a functional.

The data of *satellite-to-satellite tracking* are time changes of the distance s between two satellites (Fig. 26.1). Such a range rate ds/dt is again measured by means of the doppler principle. At present one generally uses one high and one low satellite, but the use of two low satellites which are close to each other is also possible. Considerations analogous to the preceding ones make it obvious that ds/dt has again the form (26-1), the vector X comprising now the $6 + 6$ elements of the two reference orbits.

Satellite gradiometry is designed to measure elements (or linear combinations of elements) of the second-order gradient tensor

$$\begin{bmatrix} V_{xx} & V_{xy} & V_{xz} \\ V_{xy} & V_{yy} & V_{yz} \\ V_{xz} & V_{yz} & V_{zz} \end{bmatrix}, \quad (26-23)$$

which is a symmetric matrix formed by the second derivatives of the potential V with respect to the coordinates xyz . Any second-order gradient, say V_{xz} , depends on position:

$$V_{xz} = V_{xz}(x, y, z). \quad (26-24)$$

It has, therefore, the form (26-1), with

$$X = \begin{bmatrix} x \\ y \\ z \end{bmatrix}; \quad (26-25)$$

the prescription for computing the functional F in the present case consists in differentiating V with respect to x and z and taking V_{xz} at the point with coordinates (26-25).

Very-long-baseline interferometry measures the delay τ , with which a radio signal emitted from an extragalactic radio source is received at two different places P and Q (Groten, 1979, p.50). By multiplying τ with the light velocity c we get the projection (which is a scalar product)

$$D = \vec{PQ} \cdot \vec{e} \quad (26-26)$$

of the vector \vec{PQ} connecting the two points onto the direction (supposed known) to the radio source represented by the unit vector \mathbf{e} (Fig. 26.2).

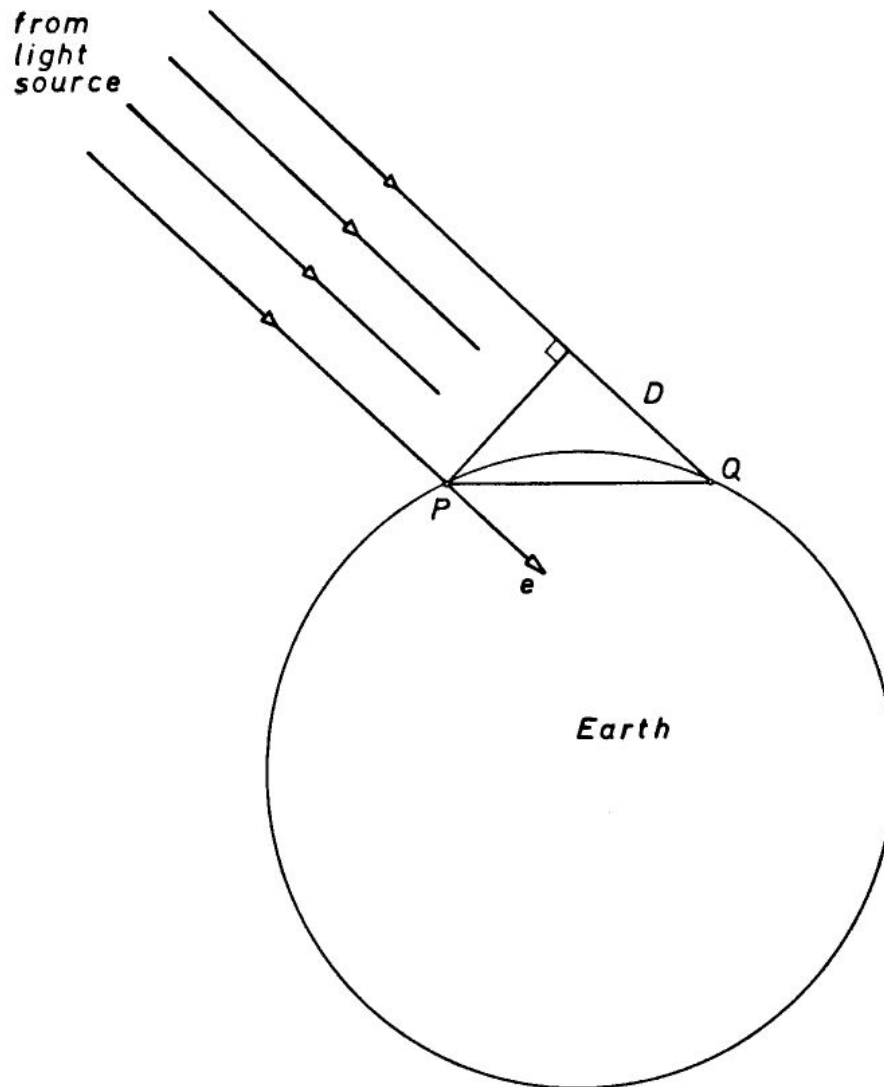


FIGURE 26.2. Very-long-baseline interferometry.

Similarly to a distance measurement (26-16), D does not depend on the gravity field, and we have a special case of (26-1) with X being given by (26-17) and with no explicit dependence on V .

These examples should make it obvious that *all* geodetic measurements, without exception, can be represented as functionals (26-1) or (26-5). This simple and general fact will be basic for the considerations to follow.

It is clear that we have taken into account only the geometrical and gravitational structure of the problem. We have abstracted from random and systematic errors, nongravitational effects, etc. Random errors will be considered later in this book, and systematic effects are assumed to have been removed by appropriate corrections. If necessary, systematic parameters can be included in the vector X in (26-1) or (26-5).

27. LINEARIZATION

Every observation l gives an equation of type (26-1) or (26-5). We thus obtain a system of functional equations

$$\begin{aligned} l_1 &= F_1(X, W) , \\ l_2 &= F_2(X, W) , \\ &\vdots \\ l_q &= F_q(X, W) , \end{aligned} \tag{27-1}$$

which are to be solved for the unknown parameters X and the potential function W .

Since the functionals F_1, F_2, \dots, F_q are non-linear, the system (27-1) is very difficult to handle directly. The usual procedure with difficult nonlinear problems is to linearize them by Taylor's theorem.

Let us introduce an approximate value X_0 for the vector X and an approximation U to the gravity potential W . The function U is called the *normal potential*; it is generally taken to be the external gravity potential of an equipotential ellipsoid (sec.2).

We put

$$X = X_0 + \delta X , \tag{27-2}$$

$$W = U + T , \tag{27-3}$$

where the differences $\delta X = X - X_0$ and $T = W - U$ are considered to be small; T is called the *anomalous potential* as usual.

Thus (26-5) becomes

$$l = F(X_0 + \delta X, U + T) \quad (27-4)$$

and a Taylor expansion gives

$$l = F(X_0, U) + a^T \delta X + LT \quad (27-5)$$

plus higher order terms, which we neglect. Here a is the column vector of ordinary partial derivatives

$$a_k = \frac{\partial F}{\partial X_k} (X_0, U) \quad (27-6)$$

of F with respect to the component X_k of the parameter vector X , taken for the approximate values X_0 and U ; a^T is the corresponding row vector, so that $a^T \delta X$ is a scalar product. The term LT is less elementary: it expresses a linear operator L acting on the function T ; its meaning will be clear from the examples to follow.

By means of the substitution

$$\delta l = F(X, W) - F(X_0, U), \quad (27-7)$$

the nonlinear system (27-1) thus becomes the linear system

$$\begin{aligned} \delta l_1 &= a_1^T \delta X + L_1 T, \\ \delta l_2 &= a_2^T \delta X + L_2 T, \\ &\vdots \\ \delta l_q &= a_q^T \delta X + L_q T. \end{aligned} \quad (27-8)$$

The linearization process will be illustrated by means of some basic special cases.

Astronomical and gravimetric observations. The equations to be linearized are (26-9), (26-10), and (26-11). Here we are considerably helped by the fact that these expressions are just ordinary functions of W_x, W_y, W_z ; X is simply the coordinate vector (26-12).

Therefore, we first linearize the gradient vector (26-7). Using index notation, we write $x = x_1$, $y = x_2$, $z = x_3$ and

$$\text{grad } W = \begin{bmatrix} W_x \\ W_y \\ W_z \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix} . \quad (27-9)$$

Thus

$$W_i = \frac{\partial W}{\partial x_i} \quad (i = 1, 2, 3) . \quad (27-10)$$

The derivatives are taken at the original point with coordinates

$$X = [x_k] \quad (k = 1, 2, 3) . \quad (27-11)$$

The approximation point is

$$X_0 = [x_k^0] ; \quad (27-12)$$

in obvious notation,

$$x_k = x_k^0 + \delta x_k . \quad (27-13)$$

Then (27-10) becomes

$$W_i = \frac{\partial W}{\partial x_i} = \left(\frac{\partial W}{\partial x_i} \right)_0 + \left(\frac{\partial^2 W}{\partial x_i \partial x_j} \right)_0 \delta x_j , \quad (27-14)$$

using the summation convention (summation over the repeated index j). The notation $()_0$ indicates that the respective quantity is to be taken at the approximation point (27-12); W_i is, of course, considered at the original point X .

We now introduce $W = U + T$ and obtain

$$\frac{\partial W}{\partial x_i} = \left(\frac{\partial U}{\partial x_i} \right)_0 + \left(\frac{\partial T}{\partial x_i} \right)_0 + \left(\frac{\partial^2 U}{\partial x_i \partial x_j} \right)_0 \delta x_j + \left(\frac{\partial^2 T}{\partial x_i \partial x_j} \right)_0 \delta x_j . \quad (27-15)$$

The last term is already of second order (T and δx_j are first-order quantities) and will be neglected. We further put

$$\left(\frac{\partial U}{\partial x_i} \right)_0 = U_i , \quad (27-16)$$

$$\left(\frac{\partial T}{\partial x_i} \right)_0 = T_i , \quad (27-17)$$

$$\left(\frac{\partial^2 U}{\partial x_i \partial x_j} \right)_0 = M_{ij} . \quad (27-18)$$

Thus (27-15) becomes finally

$$W_i = U_i + T_i + M_{ij} \delta x_j , \quad (27-19)$$

completing the linearization of the gravity vector (27-9).

The straightforward way to linearize equations (26-9) to (26-11) is to substitute (27-19) into these equations and to expand the functions in the usual way by Taylor's theorem, considering the fact that the second and third term on the right-hand side of (27-19) are small. This is simple but laborious; more efficient is an indirect procedure.

We combine (26-6) and (26-7) into the equation system

$$\begin{aligned} W_1 &= -g \cos \phi \cos \lambda , \\ W_2 &= -g \cos \phi \sin \lambda , \\ W_3 &= -g \sin \phi ; \end{aligned} \quad (27-20)$$

here all quantities refer to the original point X .

In an analogous way we write

$$\begin{aligned} U_1 &= -\gamma \cos \phi \cos \lambda , \\ U_2 &= -\gamma \cos \phi \sin \lambda , \\ U_3 &= -\gamma \sin \phi . \end{aligned} \quad (27-21)$$

Here all quantities refer to the approximation point: γ is normal gravity, and ϕ and λ are normal latitude and longitude. Cf. (Heiskanen and Moritz, 1967, p.315), where these normal geographical coordinates have been denoted by ϕ^* and λ^* .

We put

$$\begin{aligned}\phi &= \phi + \delta\phi, \\ \lambda &= \lambda + \delta\lambda, \\ g &= \gamma + \delta g,\end{aligned}\tag{27-22}$$

substitute into (27-20) and expand by Taylor. The result is readily found to be

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} + Q \begin{bmatrix} \gamma\delta\phi \\ \gamma\cos\phi\delta\lambda \\ \delta g \end{bmatrix},\tag{27-23}$$

where the matrix

$$Q = \begin{bmatrix} \sin\phi\cos\lambda & \sin\lambda & -\cos\phi\cos\lambda \\ \sin\phi\sin\lambda & -\cos\lambda & -\cos\phi\sin\lambda \\ -\cos\phi & 0 & -\sin\phi \end{bmatrix}\tag{27-24}$$

is obtained by differentiation of (27-21).

On the other hand we have (27-19), which may be written in the form

$$\begin{bmatrix} W_1 \\ W_2 \\ W_3 \end{bmatrix} = \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} + \text{grad } T + M\delta X.\tag{27-25}$$

The matrix Q is easily seen to be orthogonal (why?); therefore its inverse is simply the transpose:

$$Q^{-1} = Q^T.\tag{27-26}$$

Therefore, the comparison of (27-23) and (27-25) gives

$$\begin{bmatrix} \gamma\delta\phi \\ \gamma\cos\phi\delta\lambda \\ \delta g \end{bmatrix} = Q^T M\delta X + Q^T \text{grad } T,\tag{27-27}$$

which completes the linearization of astronomical latitude ϕ and longitude λ and of measured gravity g .

It is evident that (27-27) is, indeed, a linear function of the components δx , δy , δz of the vector δX . As regards $Q^T \text{grad } T$, it gives for each difference $\delta\phi$, $\delta\Lambda$, δg a linear expression of the form

$$\alpha_1 \frac{\partial T}{\partial x} + \alpha_2 \frac{\partial T}{\partial y} + \alpha_3 \frac{\partial T}{\partial z} = LT. \quad (27-28)$$

The operation expressed by the functional L consists in forming the partial derivatives and taking a linear combination of them. Since differentiation is a linear operation, L is indeed a linear functional.

Direction and distance measurements. The straightforward approach is to differentiate equations (26-14), (26-15) and (26-16), as outlined in (Heiskanen and Moritz, 1967, pp.220-221). The result will be differential formulas of form of eq. (5-83), *ibid*. The actual work is, however, quite cumbersome though not difficult.

Again, an indirect approach might be preferable. We put

$$\begin{bmatrix} s \sin \zeta \cos \alpha \\ s \sin \zeta \sin \alpha \\ s \cos \zeta \end{bmatrix} = \begin{bmatrix} u \\ v \\ w \end{bmatrix} = Y. \quad (27-29)$$

Then the vector Y so defined is related to the difference vector

$$\Delta X = \begin{bmatrix} x_Q - x_P \\ y_Q - y_P \\ z_Q - z_P \end{bmatrix} \quad (27-30)$$

by the linear transformation

$$Y = R \Delta X, \quad (27-31)$$

where R is the orthogonal matrix

$$R = \begin{bmatrix} -\sin \phi \cos \Lambda & -\sin \phi \sin \Lambda & \cos \phi \\ -\sin \Lambda & \cos \Lambda & 0 \\ \cos \phi \cos \Lambda & \cos \phi \sin \Lambda & \sin \phi \end{bmatrix}. \quad (27-32)$$

This is clear because u, v, w can be interpreted as rectangular coordinates in a local system in which the w -axis has the upward direction of the gravity vector and the axes u and v point north and east; the matrix R is formed by the components in the xyz system, of the unit vectors

e' , e'' , n corresponding to the uvw coordinate axes (Heiskanen and Moritz, 1967, pp.218-219).

The differentiation of (27-29) gives, in analogy to (27-23)

$$\delta Y = S \begin{bmatrix} s' \delta \zeta \\ s' \sin \zeta' \delta \alpha \\ \delta s \end{bmatrix} \quad (27-33)$$

where S is the orthogonal matrix

$$S = \begin{bmatrix} \cos \zeta' \cos \alpha' & -\sin \alpha' & \sin \zeta' \cos \alpha' \\ \cos \zeta' \sin \alpha' & \cos \alpha' & \sin \zeta' \sin \alpha' \\ -\sin \zeta' & 0 & \cos \zeta' \end{bmatrix} \quad (27-34)$$

Here we have designed by α' , ζ' , s' the "normal" equivalents of the observables α , ζ , s , so that

$$\begin{aligned} \alpha &= \alpha' + \delta \alpha, \\ \zeta &= \zeta' + \delta \zeta, \\ s &= s' + \delta s. \end{aligned} \quad (27-35)$$

The quantities α' , ζ' , s' can be computed from (26-14), (26-15), and (26-16) by using approximate coordinates X_0 and replacing ϕ, λ by ϕ, λ .

By differentiation of (27-31), on the other hand, we find

$$Y = R \delta \Delta X + \delta R \Delta X. \quad (27-36)$$

The combination of (27-33) and (27-36) gives, in view of the orthogonality of the matrix S ,

$$\begin{bmatrix} s' \delta \zeta \\ s' \sin \zeta' \delta \alpha \\ \delta s \end{bmatrix} = S^T R \delta \Delta X + S^T \delta R \Delta X. \quad (27-37)$$

The second term on the right-hand side is easily found in an indirect way. The matrix R , by (27-32), depends on ϕ and λ ; therefore δR will be a linear function of $\delta \phi$ and $\delta \lambda$. The term $S^T \delta R \Delta X$ represents, therefore, the effect of $\delta \phi$ and $\delta \lambda$ on $\delta \zeta$ and $\delta \alpha$ (there is, evidently, no effect on δs because s is independent of the gravity field); this is nothing else but the well-known effect of the deflection of the vertical on azimuth α and zenith distance ζ .

The effect on the zenith distance ζ is

$$\partial \zeta = -\xi \cos \alpha' - \eta \sin \alpha' \quad (27-38)$$

and on the azimuth α ,

$$\partial \alpha = \xi \sin \alpha' \cot \zeta' + \eta (\tan \phi - \cos \alpha' \cot \zeta') . \quad (27-39)$$

Cf. (18-11) and (18-13); we have replaced α and ζ by α' and ζ' in agreement with our present notation. We have used the symbols $\partial \zeta$ and $\partial \alpha$ to indicate the partial influence, on ζ and α , of the changes $\delta \phi$ and $\delta \Lambda$, which are related to the deflection components ξ and η by

$$\xi = \delta \phi , \quad \eta = \delta \Lambda \cos \phi . \quad (27-40)$$

The comparison of (27-24) and (27-32) shows that, for $\phi = \phi$ and $\Lambda = \lambda$,

$$R = -Q^T ; \quad (27-41)$$

the geometrical interpretation of this fact is left to the reader.

In view of the relations (27-38) to (27-41), eq. (27-37) takes the final form

$$\begin{bmatrix} s' \delta \zeta \\ s' \sin \zeta' \delta \alpha \\ \delta s \end{bmatrix} = -S^T Q^T \begin{bmatrix} \delta x_Q - \delta x_P \\ \delta y_Q - \delta y_P \\ \delta z_Q - \delta z_P \end{bmatrix} + K \begin{bmatrix} \delta \phi \\ \cos \phi \delta \Lambda \end{bmatrix} \quad (27-42)$$

where

$$K = \begin{bmatrix} -s' \cos \alpha' & -s' \sin \alpha' \\ s' \sin \alpha' \cos \zeta' & s' (\tan \phi \sin \zeta' - \cos \alpha' \cos \zeta') \\ 0 & 0 \end{bmatrix} ; \quad (27-43)$$

the matrices Q and S are given by (27-24) and (27-34).

These examples will illustrate how the linearized equations (27-8) can be obtained. Similar linearizations can be found in (Eeg and Krarup, 1975) and (Grafarend, 1977, 1978).

The reader will have noticed the close relation of the present linearization to sec. 18: we are looking at the same subject from a more general point of view.

28. VARIATIONAL PRINCIPLES

Let us take up the linearized system (27-8). To simplify the notation, we replace

$$\delta l_i \text{ by } l_i, \quad \delta X \text{ by } X,$$

obtaining

$$\begin{aligned} l_1 &= a_1^T X + L_1 T, \\ l_2 &= a_2^T X + L_2 T, \\ &\vdots \\ l_q &= a_q^T X + L_q T. \end{aligned} \tag{28-1}$$

where X is a p -vector (a $p \times 1$ matrix). We finally put

$$l = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_q \end{bmatrix}, \quad A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_q^T \end{bmatrix}, \tag{28-2}$$

and

$$B = \begin{bmatrix} L_1 \\ L_2 \\ \vdots \\ L_q \end{bmatrix}. \tag{28-3}$$

Here l is a q -vector (a $q \times 1$ matrix), A is a $q \times p$ matrix, and B is a linear operator, formed of the q linear functionals L_k . We assume $p < q$.

With these notations, the system (28-1) becomes

$$l = AX + BT. \tag{28-4}$$

These equations hold exactly (within the limits of linearization) if there are no measuring errors. Because of these errors, the quantity $l - AX - BT$ will not be exactly zero; let us put

$$l - AX - BT = n, \quad (28-5)$$

so that n is the effect of measuring errors on the observation vector l . Writing this equation in the form

$$l = AX + BT + n, \quad (28-6)$$

we see that we have recovered the basic observation equation (16-1) for least-squares collocation. We have, however, derived it from a quite general point of view, and we shall also continue to treat the problem more generally.

Improperly posed problems. A problem is called *properly posed* if the solution satisfies the following three requirements:

- (1) existence,
- (2) uniqueness,
- (3) stability.

This means that a solution must exist for arbitrary (within a certain range) data, that there must be only one solution, and that this solution must depend continuously on the data. If one or more of these requirements are violated, then we have an *improperly posed*, or *ill-posed*, problem.

For a long time it was thought that only properly posed problems are physically meaningful. In fact, deterministic processes, as considered in classical mechanics, depend uniquely and continuously on the initial data--this is the essence of causality--and thus correspond to properly posed problems.

Only relatively recently it was recognized that there are important problems that are not properly posed. There is now an extensive literature on improperly posed problems; we mention only two easily accessible books: (Lavrentiev, 1967) and, especially, (Tikhonov and Arsenin, 1977), and the review article (Nashed, 1974). Geodetic applications are considered in (Schwarz, 1978b); for instance, the downward continuation of gravity is an ill-posed problem. The relation between least-squares collocation and improperly posed problems was pointed out by Neyman (1975, 1977).

Our present task, the determination of the earth's gravitational field from measurements, is a typical improperly posed problem. The potential is so irregular that it cannot be completely described by any finite set of

parameters; on the other hand, we have only a finite number of measurements. Hence, there is no unique solution, and Condition 2 is violated.

We shall try to approach the present geodetic problem from the point of view of the theory of improperly posed problems. Let us put

$$z = \begin{bmatrix} X \\ T \end{bmatrix} . \quad (28-7)$$

Since $X \in R^P$ and $T \in H$, where H is some space of harmonic functions, the symbol z denotes an element of the product space $R^P \times H$; cf. the footnote on p.221. We further put

$$G = [A \quad B] , \quad (28-8)$$

so that (28-4) becomes

$$Gz = 1 . \quad (28-9)$$

Since $1 \in R^q$, the linear operator G denotes a mapping

$$G : R^P \times H \rightarrow R^q . \quad (28-10)$$

The solution, if it exists, may be written formally in the form

$$z = G^{-1}1 \quad (28-11)$$

but it will certainly not be unique. Therefore, G^{-1} is not an inverse operator in the usual sense; it has the character of a generalized inverse operator, analogous to generalized matrix inverses; cf. sec. 21.

A standard method for solving improperly posed problems is *Tikhonov regularization*. It consists in minimizing the nonlinear functional

$$M^\alpha[z, 1] = \|Gz - 1\|^2 + \alpha \Omega(z) , \quad (28-12)$$

where α is a numerical parameter and $\Omega(z)$ is a so-called *stabilizing functional* (Tikhonov and Arsenin, 1977, pp.51,57), which may be taken as the square of some norm,

$$\Omega(z) = \|z\|^2 \quad (28-13)$$

(*ibid.*, p.72). In this way, a unique solution can usually be obtained.

Using (28-13) and very slightly generalizing (28-12) by the introduction of a second numerical parameter β , we get the condition

$$\alpha \|z\|^2 + \beta \|Gz - 1\|^2 = \text{minimum} . \quad (28-14)$$

By (28-6),

$$Gz - 1 = AX + BT - 1 = -n \quad (28-15)$$

is the (negative) error of satisfying (28-4), so that (28-14) may also be written

$$\alpha \|z\|^2 + \beta \|n\|^2 = \text{minimum} . \quad (28-16)$$

This condition means minimizing a weighted square average of the "function norm" $\|z\|$ and the "error norm" $\|n\|$. The desired result will be a solution of (28-6) subject to the condition (28-16).

Choice of the norms. The choice of the error norm $\|n\|$ is straightforward, since $n \in R^q$. Any regular quadratic norm in q -dimensional Euclidean space R^q can be written

$$\|n\|^2 = n^T Q n \quad (28-17)$$

with a positive definite regular symmetric $q \times q$ "weight matrix" Q . Denoting its inverse by D , we may also write

$$\|n\|^2 = n^T D^{-1} n . \quad (28-18)$$

If the elements of the vector n are random quantities in a statistical sense, then D may be regarded as the covariance matrix of the noise n ; otherwise (28-17) is just a metric in a purely geometrical sense.

A quadratic norm $\|X\|$ for the parameter vector X is similarly

$$\|X\|^2 = X^T P X , \quad (28-19)$$

with a regular positive definite symmetric weight matrix P .

Finally, the norm for T will be selected as a norm in a Hilbert space with a kernel function $K(P,Q)$, which is the equivalent of quadratic norms such as (28-17) and (28-19) in an infinitely-dimensional function space (sec.24); hence

$$\|T\|^2 = (T,T) . \quad (28-20)$$

If we assume X and T independent of each other, then the norm of

$$z = \begin{bmatrix} X \\ T \end{bmatrix}$$

is simply

$$\|z\|^2 = \|X\|^2 + \|T\|^2 , \quad (28-21)$$

and the Tikhonov condition (28-16) becomes

$$\alpha(\|X\|^2 + \|T\|^2) + \beta \|n\|^2 = \text{minimum} \quad (28-22)$$

or

$$\alpha(T,T) + \alpha X^T P X + \beta n^T Q n = \text{minimum} . \quad (28-23)$$

The matrix P imposes a restriction on the variability of the vector X . In a statistical interpretation, P^{-1} is an *a priori* variance matrix for X . If the variance of a random quantity is small, then this quantity can vary only within narrow limits. The larger the variance, the larger variations are possible; and if the variance goes to infinity, the variation of our quantity becomes completely free.

If the parameter vector is allowed to vary freely without restriction, then, in statistical terminology, each component of X should have infinite variance or zero weight, which means $P = 0$. In this way parameters are usually treated in least-squares adjustment, and we have also treated them so in Part B, sections 16 *et seq.* Then $\|X\| = 0$ and (28-23) reduces to

$$\alpha(T,T) + \beta n^T D^{-1} n = \text{minimum} . \quad (28-24)$$

In the following section we shall use the minimum condition (28-24); the general condition (28-23) will be briefly considered in sec. 30.

Maximum and minimum problems involving unknown functions are called variational problems. Therefore, (28-23) and (28-24) may be considered geodetic variational principles.

29. SOLUTION OF A VARIATIONAL PROBLEM

In the preceding section we have seen that the determination of the parameter p-vector X and of the potential T from the q-vector l of observations ($p < q$) can be reduced to the solution of the linear system

$$AX + BT + n = l \quad (29-1)$$

subject to the variational principle (28-24),

$$\alpha(T, T) + \beta n^T D^{-1} n = \text{minimum} . \quad (29-2)$$

Pure collocation. As a preparation, let us assume errorless observations. Then $n = 0$ and (29-2) reduces to

$$(T, T) = \text{minimum} \quad (29-3)$$

(we have put $\alpha = 1$ without loss of generality). We furthermore assume $X = 0$ (no systematic effects). Then (29-1) becomes

$$BT = l \quad (29-4)$$

where l is given. The desired T is that function T , satisfying (29-4), which minimizes (29-3).

We solve the problem by means of a Lagrange multiplier. Instead of minimizing (29-3) under the side condition (29-4), we form the unconditional minimum of the function

$$\Phi = \frac{1}{2} (T, T) - k^T (BT - l) , \quad (29-5)$$

where the q-vector k serves as a Lagrange multiplier.

A necessary condition for a minimum is the vanishing of the differential of Φ :

$$d\phi = (T, dT) - k^T B dT = 0 . \quad (29-6)$$

Note that we have formed this differential as if T were a vector. Let us point out, however, that dT is not an ordinary differential of T , but what is called, in the calculus of variations, a first variation, that is, a change in the function T . In fact, (29-6) is the *Euler equation* corresponding to the variational problem (29-3). It may be found as follows. In (29-5) we replace T by $T + \epsilon\tau$, where ϵ is a small parameter:

$$\begin{aligned} \phi_\epsilon &= \frac{1}{2}(T + \epsilon\tau, T + \epsilon\tau) - k^T(BT + \epsilon B\tau - 1) = \\ &= \frac{1}{2}(T, T) - k^T(BT - 1) + \frac{1}{2}\epsilon(T, \tau) + \frac{1}{2}\epsilon(\tau, T) - \\ &\quad - \epsilon k^T B\tau + \frac{1}{2}\epsilon^2(\tau, \tau) . \end{aligned}$$

By symmetry, $(\tau, T) = (T, \tau)$. We then subtract (29-5) and divide by ϵ . On letting $\epsilon \rightarrow 0$, we thus get

$$\lim_{\epsilon \rightarrow 0} \frac{\phi_\epsilon - \phi}{\epsilon} = (T, \tau) - k^T B\tau = 0 , \quad (29-7)$$

which is (29-6), with $dT = \epsilon\tau$.

The function dT in (29-6) is completely arbitrary; it need not even be small since a numerical factor does not matter; of course, dT must belong to the Hilbert space under consideration.

By the reproducing property (24-2) there is

$$dT(Q) = (dT(P), K(P, Q)) \quad (29-8)$$

or briefly,

$$dT = (dT, K) = (K, dT) . \quad (29-9)$$

Hence,

$$BdT = (BK, dT) , \quad (29-10)$$

and (29-6) becomes

$$(T, dT) - (k^T BK, dT) = 0$$

or

$$(T - k^T B K, dT) = 0 . \quad (29-11)$$

since dT is arbitrary, there must be

$$T - k^T B K = 0$$

or

$$T = k^T B K . \quad (29-12)$$

This is an important result. What does it mean? In view of (28-3) this is nothing else than

$$T(Q) = \sum_{i=1}^q k_i L_i^P K(P, Q) ; \quad (29-13)$$

L_i^P means the operator L_i applied to the variable P . Now, (29-13) is identical to (25-3) and (25-4), with P and Q interchanged and $b_i = k_i$. Thus, the best approximation for $T(Q)$ is, in fact, a linear combination of the base functions (25-3)!

The rest is straightforward. Considering T a scalar, we may transpose (29-12):

$$T = (BK)^T k , \quad (29-14)$$

and substitute into (29-4):

$$B(BK)^T k = 1 . \quad (29-15)$$

The $q \times q$ matrix

$$C = B(BK)^T \quad (29-16)$$

has, by (25-18) and (11-14), the elements

$$C_{ij} = L_i^P L_j^Q K(P, Q) . \quad (29-17)$$

Now (29-15) may be solved for k :

$$k = C^{-1}l , \quad (29-18)$$

so that (29-14) becomes

$$T = (BK)^T C^{-1}l , \quad (29-19)$$

which is identical to (25-17).

We thus have obtained (25-17) as a consequence of the minimum norm principle (29-3). This is neither the shortest nor the most complete proof since the condition (29-6) is only necessary but not sufficient; the advantage of the present derivation is the treatment as a straightforward solution of a variational principle by standard techniques (Euler equation). Furthermore, it will essentially simplify the treatment of the general case.

The case $\alpha = \beta = 1$. As a second step, let us consider the general equation (29-1), but put, in the Tikhonov condition (29-2), $\alpha = \beta = 1$, so that

$$(T, T) + n^T D^{-1} n = \text{minimum} , \quad (29-20)$$

to be solved under the side condition (29-1). We thus have to form the unconditional minimum of the function

$$\phi = \frac{1}{2}(T, T) + \frac{1}{2} n^T D^{-1} n - k^T (AX + BT + n - l) \quad (29-21)$$

The differential is

$$d\phi = (T, dT) + n^T D^{-1} dn - k^T (AdX + BdT + dn) , \quad (29-22)$$

where dX and dn are ordinary vector differentials. On rearranging we get, using (29-10),

$$d\phi = (T - k^T BK, dT) + (n^T D^{-1} - k^T) dn - k^T AdX = 0 . \quad (29-23)$$

Since dT , dn , and dX are arbitrary, $d\phi = 0$ can only hold if

$$T - k^T BK = 0 , \quad (29-24)$$

$$n^T D^{-1} - k^T = 0 , \quad (29-25)$$

$$k^T A = 0 . \quad (29-26)$$

The first equation gives

$$T = k^T B K , \quad (29-27)$$

identical to (29-12) or (29-13). Again, the solution is a linear combination of base functions (25-3)!

The transposition of (29-27) gives

$$T = (BK)^T k , \quad (29-28)$$

so that

$$BT = B(BK)^T k = Ck , \quad (29-29)$$

using the abbreviation (29-16). Eq. (29-25) gives

$$n^T = k^T D \quad \text{or} \quad n = Dk , \quad (29-30)$$

Eq. (29-1) may be written

$$1 - AX = BT + n , \quad (29-31)$$

and substituting (29-29) and (29-30) we get

$$1 - AX = (C + D)k , \quad (29-32)$$

so that

$$k = (C + D)^{-1}(1 - AX) . \quad (29-33)$$

We substitute this into (29-26), transposed as

$$A^T k = 0 ,$$

obtaining

$$A^T(C + D)^{-1}1 - A^T(C + D)^{-1}AX = 0 ,$$

so that

$$X = [A^T(C + D)^{-1}A]^{-1}A^T(C + D)^{-1}1 , \quad (29-34)$$

which determines the parameter vector X . The substitution of (29-33) into (29-28) then gives the potential:

$$T = (BK)^T(C + D)^{-1}(1 - AX) . \quad (29-35)$$

The general case. Take finally the general Tikhonov condition (29-2)

$$\alpha(T, T) + \beta n^T D^{-1}n = \text{minimum} \quad (29-36)$$

for solving the equation (29-1)

$$AX + BT + n = 1 . \quad (29-37)$$

The condition (29-36) is, for $\alpha \neq 0$, equivalent to

$$(T, T) + n^T \left(\frac{\alpha}{\beta} D\right)^{-1}n = \text{minimum} , \quad (29-38)$$

so that we only have to replace, in (29-34) and (29-35), D by $\alpha D/\beta$. This gives

$$X = [A^T(\beta C + \alpha D)^{-1}A]^{-1}A^T(\beta C + \alpha D)^{-1}1 , \quad (29-39)$$

$$T = (\beta BK)^T(\beta C + \alpha D)^{-1}(1 - AX) , \quad (29-40)$$

which is the solution of (29-37) under the general Tikhonov condition (29-36).

Obviously, to various ratios $\alpha : \beta$ there corresponds a different weighting between the square of the "function norm", (T, T) , and of the "error norm", $n^T D^{-1}n$. Pure collocation, with the condition (29-3) and $n = 0$, fits the solution exactly to the data. This is unsuited for real data in the presence of measuring errors, because then the solution is distorted by faithfully reproducing all measuring errors; we risk to get spurious oscillations.

Therefore, some balance between α and β in (29-36) must be found. But how? We might use some trial-and-error procedure, but this does not seem very satisfactory. A theoretically motivated, in a certain sense optimal, solution to this problem is found by statistical considerations, as we shall see in the following section.

We also mention that the result for the "errorless" condition (29-3) cannot be obtained by simply putting $\alpha = 1$, $\beta = 0$, as might be expected at first sight. The essential feature with (29-3) is that $n = 0$ and $D = 0$. We, therefore, have to put $D = 0$ in (29-39) and (29-40). As a consequence, β cancels then, and we obtain

$$X = (A^T C^{-1} A)^{-1} A^T C^{-1} 1, \quad (29-41)$$

$$T = (BK)^T C^{-1} (1 - AX). \quad (29-42)$$

For $A = 0$ (no systematic parameters), the last equation reduces to (29-19) as it should.

A final word on the solution of these variational principles by an Euler equation. Any Euler equation gives only a necessary, not a sufficient, condition for a minimum. It is not too difficult to show that our solutions do indeed give a minimum. The proof goes along well-known lines, by an extension of the reasoning on pp. 119-121 and pp. 207-209.

30. LEAST-SQUARES COLLOCATION AND RELATED MODELS

The solution of the variational problem defined by (29-1) and (29-2) for $\alpha = \beta = 1$:

$$AX + BT + n = 1, \quad (30-1)$$

$$(T, T) + n^T D^{-1} n = \text{minimum} \quad (30-2)$$

has been found to be given by (29-34) and (29-35). On substituting

$$C + D = C_{tt} + C_{nn} = \bar{C} \quad (30-3)$$

we may write (29-35) as

$$T = (BK)^T \bar{C}^{-1} (1 - AX). \quad (30-4)$$

The substitution (30-3) is motivated as follows. In agreement with (14-5) we put

$$BT = t. \quad (30-5)$$

In sec. 25 we have had $1 = BT = t$ since there was $n = 0$ and $x = 0$. Hence (25-19) and (29-16) give

$$B(BK)^T = C_{tt} = C \quad (30-6)$$

which, together with (14-36) and (16-29) explains (30-3).

Any signal s_k is a linear functional of T in agreement with (25-43):

$$s_k = S_k T; \quad (30-7)$$

for the m -vector $s = [s_k]$ we may write

$$s = ST, \quad (30-8)$$

where

$$S = \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_m \end{bmatrix} \quad (30-9)$$

is a linear operator $S : H \rightarrow R^m$, since T is an element of a Hilbert space H of harmonic functions and $s \in R^m$.

The application of S to (30-4) gives

$$s = S(BK)^T \bar{C}^{-1} (1 - AX). \quad (30-10)$$

Now $S(BK)^T$ is simply the matrix with elements (14-30):

$$S(BK)^T = C_{st} \quad (30-11)$$

so that, together with (29-34), we get

$$X = (A^T \bar{C}^{-1} A)^{-1} A^T \bar{C}^{-1} l, \quad (30-12)$$

$$s = C_{nt} \bar{C}^{-1} (l - AX). \quad (30-13)$$

Noting that X and s are estimates \hat{X} and \hat{s} , we see that these two equations are identical to the basic least-squares collocation equations (16-36) and (16-37). In sec. 16 we have derived these equations from the finite minimum principle

$$s^T \bar{C}^{-1} s + n^T D^{-1} n = \text{minimum}, \quad (30-14)$$

s being the *finite* vector (16-11), whereas in sec. 29, T has been treated as an element in Hilbert space, the condition (30-2) being a variational principle for the unknown *function* T .

It is in order now to compare the two approaches. The approach of sec. 16 was elementary in the sense that only finite matrix operations were used. It was rigorous since no approximations were involved, but it was not fully satisfying as, for instance, the discussions on pp.117-119 show. In fact, what essentially we look for is the full anomalous gravity field, that is, for the *function* T , and T is an element of an infinitely-dimensional function space, of a Hilbert space.

Thus, even in finite matrix formulas such as (30-6) and (30-7) (letting the vector s comprise only those finitely many signals we are computing), Hilbert space lurks invisibly in the background; to use another metaphor, these formulas represent the finite-dimensional surface of infinitely-dimensional Hilbert space.

The treatment of sec. 16 was, in a sense, an acrobatic effort to (almost completely) avoid Hilbert space, working only with finite matrices. It necessarily stayed at the surface of the problem, and a deeper understanding requires the genuine use of Hilbert space.

In particular, the present Hilbert space theory permits an interpretation of least-squares collocation as a least-squares adjustment in an infinitely-dimensional space. In fact, put

$$v = \begin{bmatrix} T \\ n \end{bmatrix}. \quad (30-15)$$

Since $T \in H$ and $n \in R^q$, we see that v is an element of the product space $H \times R^q$ (see the footnote on p.222), which is also a Hilbert space with norm

$$\|v\|^2 = \|T\|^2 + \|n\|^2 = (T, T) + n^T D^{-1} n . \quad (30-16)$$

By means of the operator $M : H \times R^q \rightarrow R^q$, defined by

$$M = [B \quad I] \quad (30-17)$$

where I is the $q \times q$ unit matrix, we may write (30-1) in the form

$$AX + Mv = 1 , \quad (30-18)$$

to be solved under the least-squares condition

$$\|v\|^2 = \text{minimum} . \quad (30-19)$$

If v were a simple vector, then (30-18) together with (30-19) would be ordinary least-squares adjustment by condition equations with parameters; now we have an infinitely-dimensional analogue. Cf. also (Krarup, 1969, pp.39-41), (Rummel, 1976), (Ecker, 1977), and the remarks at the end of sec. 25.

The present Hilbert space treatment also avoids the objections against an interpretation as a finite-dimensional adjustment problem mentioned on pp.117-118. In fact, the infinite vector v given by (30-15) enters fully both in the condition equations (30-18) and the least-squares condition (30-19).

It might be asked whether the infinite dimensionality of Hilbert space is essential to the problem. Operations in Hilbert space show formal similarities with operations in finite-dimensional space but are qualitatively different, just as differential equations are qualitatively different from difference equations although there are formal similarities.

It is true that the gravity field at satellite elevations can be adequately described by a spherical-harmonic expansion truncated at a sufficiently high degree; the coefficients of such a development do form a finite-dimensional vector. Thus, if we exclusively work at satellite altitudes, we might, in fact, replace Hilbert space by a finite-dimensional space, without essentially impairing the accuracy.

This situation changes essentially if we consider the gravity field at the earth's surface by including terrestrial observations such as gravity anomalies or deflections of the vertical. The detailed gravity field at the earth's surface cannot be adequately described by a spherical-harmonic expansion, neither from a theoretical point of view--because the conver-

gence cannot be guaranteed--nor from a practical point of view--because, if at all possible, such an expansion would require an excessively high number of terms, which is beyond the capacity of any present digital computer.

Hence, the general replacement of Hilbert space by a finite-dimensional space is neither theoretically adequate nor practically feasible.

Nor is it necessary since, as we have seen, the final collocation equations are finite-dimensional matrix formulas; the Hilbert space character expresses itself only in the fact that covariances are propagated, not by matrix operations, but by linear operations (such as differentiation) of covariance functions.

Geometrical Interpretation. The geometrical interpretation of the present model is quite similar as that of "pure" collocation without noise as given in sec. 25.

So far, H has been a Hilbert space of harmonic functions. It is slightly more convenient and completely equivalent to consider H as a Hilbert space of infinite sequences s given by (25-24). This is more convenient because in this way we can work with (infinite) matrices instead of operators, and it is equivalent because of the isomorphism (25-29).

Thus (30-1) becomes

$$AX + Bs + n = l. \quad (30-20)$$

The matrix A and the vectors l , X , and n are the same as before: l is a given q -vector comprising the q observations, X is a p -vector ($p < q$) consisting of p unknown parameters, n is an unknown q -vector comprising random measuring errors, and A is a given $q \times p$ matrix. Instead of T we now have the Hilbert vector s consisting, e.g., of the infinite set of spherical-harmonic coefficients of the potential T , and the given operator B is now a $q \times \infty$ matrix.

This case of least-squares collocation with parameters can easily be reduced to the corresponding case without parameters, by eliminating the p parameters from the system of q linear equations (30-20). Then we are left with a system of $q - p$ linear equations, which may be written in the form

$$Ms + Nn = y. \quad (30-21)$$

Putting

$$q - p = r, \quad (30-22)$$

which is a positive integer, we see that y is a given r -vector, M is a $r \times q$ matrix, and N is a $r \times q$ matrix.

Let v be the infinite vector comprising both noise and signal:

$$v = [n_1 \ n_2 \ \dots \ n_q \ s_1 \ s_2 \ s_3 \ \dots]^T. \quad (30-23)$$

It may be considered as a vector in the product space $R^q \times H$, the norm being given by

$$\|v\|^2 = \|s\|^2 + \|n\|^2. \quad (30-24)$$

This is the same as (30-16) in view of (25-35); note that s is now the infinite vector (25-24) and not the finite vector (16-11) entering in (30-14). Hence (30-19) is the same as

$$\|v\|^2 = \text{minimum} \quad (30-25)$$

also in the present definition of v by (30-23).

Now (30-21) may be abbreviated as

$$Rv = y. \quad (30-26)$$

This system of $r = q - p$ equations has exactly the same structure as the system (21-31),

$$Bs = 1,$$

of which the solution is given by (21-33) or (25-26), with s replaced by v , and the condition

$$\|s\|^2 = \text{minimum}$$

replaced by (30-25).

Thus we can directly take over to the present case the geometrical interpretations discussed in sec. 25: minimum norm, cf. Fig. 25.1, and minimum error norm, cf. Fig. 25.3, which correspond to the classical Gaussian conditions, least-squares and minimum variance.

The geometrical meaning of the least-squares collocation problem defined by (30-1) and (30-2) (or, equivalently, by (30-26) and (30-25)) may thus be formulated in two ways:

1. *Minimum norm.* In the product space $R^q \times H$, find the shortest distance from the origin to the hyperplane (D in Fig. 25.1) of codimension $r = q - p$, defined by the observation equations (30-1) or, equivalently, by (30-26).

2. *Minimum error norm.* In the dual space to $R^q \times H$, orthogonally project the functionals to be determined, onto the r -dimensional subspace spanned by the given functionals forming the vector y (corresponding to the subspace H'_1 in Fig. 25.3).

Statistical interpretation. As we have repeatedly pointed out, especially in sec. 12, least-squares collocation has an analytical and a statistical aspect, both of which are fundamental. The analytical structure forms the firm skeleton of the method and must, therefore, be correct. Hence we have now derived the least-squares collocation formulas in a purely analytical way, starting from the observations and using a variational principle. In a way, it was an exercise in Hilbert space geometry.

The kernel function $K(P, Q)$ and the matrix D defining the metric in the various spaces used have been arbitrary in principle. However, statistics serves as a guide for the proper choice of the function K and the matrix D by interpreting them as covariances. Therefore, we shall use the following sections for a detailed study of the statistical aspects.

Let us here only note that if $K(P, Q)$ is interpreted as the covariance function for the potential T and D as the covariance matrix for the noise n , then the parameters α and β in (29-2) must both be equal to one; cf. secs. 16 and 17.

A-priori weights for parameters. Let us now return to the variational principle (28-23). Without loss of generality we put again $\alpha = \beta = 1$ because, even if $\alpha \neq 1 \neq \beta$, we may absorb these constants in the kernel function and in the weight matrices P and Q . Then (28-23) becomes with $Q = D^{-1}$,

$$(T, T) + n^T D^{-1} n + X^T P X = \text{minimum} . \quad (30-27)$$

This variational principle can now be treated in exactly the same manner as the condition (29-20) in sec. 29. We form

$$\Phi = \frac{1}{2}(T, T) + \frac{1}{2} n^T D^{-1} n + \frac{1}{2} X^T P X - k^T (AX + BT + n - l) \quad (30-28)$$

and get for the differential

$$d\Phi = (T - k^T B K, dT) + (n^T D^{-1} - k^T) dn + (X^T P - k^T A) dX . \quad (30-29)$$

The condition $d\phi = 0$ leads to three equations, two of which are identical to (29-24) and (29-25), whereas the third becomes

$$X^T P - k^T A = 0, \quad (30-30)$$

replacing (29-26). The reasoning from (29-27) until (29-33) remains unchanged. We now write (29-33) as

$$k = \bar{C}^{-1}(1 - AX) \quad (30-31)$$

and (30-30) as

$$X = P^{-1} A^T k, \quad (30-32)$$

whence

$$X = P^{-1} A^T \bar{C}^{-1} (1 - AX).$$

This is solved for X to obtain

$$X = (A^T \bar{C}^{-1} A + P)^{-1} A^T \bar{C}^{-1} 1. \quad (30-33)$$

The matrix P may be interpreted statistically as an a-priori weight matrix for the parameter vector X . This is sometimes used in least-squares adjustment, too; cf. (Wolf, 1968, p.525). Such a technique has also been applied, e.g., in (Lerch et al., 1977).

It is immediately seen that (29-35) remains unchanged. Thus, the only effect of replacing the variational principle (30-2) by (30-27) is to replace the parameter estimation equation (30-12) by (30-33); the signal estimation equation (30-13) remains the same.

A related and even more general treatment may be found in (Dermanis, 1978). Another method related to collocation is Bjerhammar's (1975) "reflexive prediction". The paper (Wolf, 1977) contains a clear account of relations between adjustment, "discrete collocation" (with a finite-dimensional space instead of Hilbert space), and reflexive prediction. For a comparison between Bjerhammar's method and least-squares collocation cf. (Sjöberg, 1978).

Why collocation? For the operational approach to physical geodesy, which we have pursued from sec. 26, starting from the available geodetic

measurements, to the present point, is there an alternative to collocation? We have seen that all measurements are nonlinear functionals of the potential V and of certain parameters, which after linearization become linear functionals of T and X ; this leads necessarily to the linear system of functional equations (29-1). As we have pointed out on p. 85, collocation in a mathematical sense is the approximation of a function by fitting an analytical expression to q given linear functionals; for a very simple example see (Moritz and Sunkel, 1978, p.32).

Such collocation methods are used in applied mathematics for the approximate solution of differential equations, etc. There the functionals are usually supposed to be given in a mathematically exact way, and the analytical expression is required to fit these data exactly.

Such a "pure" or "mathematical" collocation is, in general, not adequately applicable to our present geodetic problem, in view of the inevitable random measuring errors (noise). An exception is, for instance, least-squares interpolation of gravity anomalies (interpolation is a special case of collocation, in which the functionals are simply the values of the function at discrete points); here, the measuring errors of gravity are considered to be negligibly small. Generally, measuring errors, or "noise", must be suitably taken into account: we have a problem of "collocation with noise". (This is true for the linearized problem, but may be said to hold even for the original nonlinear problem because measurements, by their very nature, are nonlinear functionals of V and X , and the notion of collocation may be extended, in a natural way, also to the fitting of nonlinear functionals.)

Thus the operational approach to physical geodesy, starting from the measurements, inevitably leads to a collocation problem (with noise). What can be chosen in different ways, is the analytical expression for approximating the potential. Let us, therefore, consider some alternative possibilities.

Alternative base functions. Besides kernel functions, many other functions can be used as base functions for expressions such as (25-4),

$$T(P) \doteq \sum_{i=1}^q b_i \phi_i(P) , \quad (30-34)$$

approximating the anomalous potential T . We only mention polynomials in polynomial interpolation, trigonometric functions in trigonometric interpolation and approximation and, in geodetic applications, multiquadric functions (Hardy, 1976). Particularly useful are *spline functions* and other finite elements, for an elementary discussion and comparison with kernel

functions cf. (Moritz, 1978b). An interesting geodetic application of spline functions is to a fast computation of covariance functions (Sunkel, 1978b).

These functions can be very well suited for particular applications, but they cannot be used to solve the general operational problem of physical geodesy because they are not harmonic outside the earth, which is required for approximating the potential T .

Spherical harmonic functions can be used to represent the potential, and they are, in fact, fundamental for this purpose. From a practical point of view, they are particularly suited for representing the global field at satellite altitudes. The sample functions of Giacaglia and Lundquist (1972) are finite linear combinations of spherical harmonics in a form convenient for certain purposes. For local representations of the detailed gravity field, however, spherical harmonics are not applicable. This excludes their use as base functions in the present general context.

Kernel functions can equally well be used for local and global purposes; this explains their application in our general operational approach.

Choice of norm. Let us finally return to the Tikhonov principle (28-22). Why did we use quadratic norms, arriving at (28-23)? The reason is simplicity: only then will we get a linear variational (Euler) equation, leading to a linear combination of base functions.

This motivates the use of a quadratic norm, which is the inner product of T with itself:

$$\|T\|^2 = (T, T),$$

that is, the use of Hilbert space. But why a Hilbert space with kernel functions, not some other Hilbert space?

The reason is that only then the Euler equation will, in fact, lead to a linear combination (12-11) of base functions. Let us illustrate this by means of a counterexample. If we restrict ourselves to the interpolation of functions f defined on a sphere σ , then a very obvious choice of norm would be the L_2 norm

$$\|f\|^2 = (f, f) = \iint_{\sigma} f^2 d\sigma, \quad (30-35)$$

$d\sigma$ being the surface element of the sphere. The Hilbert space so defined does not have a kernel function in the proper sense; however, it may be considered as a Hilbert space with a "generalized" kernel function, which is a Dirac delta function defined as a two-dimensional analogue of (4-22):

$$\delta(P, Q) = \delta(Q, P) ,$$

$$\delta(P, Q) = 0 \quad \text{if} \quad P \neq Q , \quad (30-36)$$

$$\iint_{\sigma} \delta(P, Q) d\sigma_P = 1 \quad \text{for fixed } Q .$$

Then

$$\begin{aligned} (f(P), K(P, Q)) &= (f(P), \delta(P, Q)) = \iint_{\sigma} f(P) \delta(P, Q) d\sigma_P = \\ &= f(Q) \iint_{\sigma} \delta(P, Q) d\sigma_P = f(Q) , \end{aligned}$$

in agreement with the definition (24-2).

For our interpolation on the sphere we thus have, by (25-3),

$$\phi_j(P) = \delta(P, P_j) . \quad (30-37)$$

These "functions" are zero outside the interpolation points. The interpolation function, the linear combination (25-4), has, therefore, the property that it is zero everywhere outside the points at which the functional values are given; there is no smooth interpolation.

This simple example will illustrate that only quadratic norms with kernel functions can be used.

In sec. 24 we have seen that a kernel function norm is a natural extension, to Hilbert space, of the usual quadratic norm

$$x^T K^{-1} x ,$$

with a positive definite regular matrix K , of vectors x in a finite-dimensional space.

Thus, the Tikhonov approach with a quadratic norm inevitably leads to collocation with kernel functions, other base functions cannot be obtained in this way.

31. STOCHASTIC PROCESSES ON THE CIRCLE

In the following sections we shall discuss statistical aspects of collocation. In particular, we shall consider some stochastic processes on the sphere which might be suited as statistical models for the earth's gravitational field; the treatment closely follows (Moritz, 1978c). As a preparation, let us first look at stochastic processes on the circle, which are considerably simpler and already show essential theoretical features.

A continuous and continuously differentiable function $f(t)$ on the unit circle $0 \leq t < 2\pi$ can be expanded into a uniformly convergent Fourier series (Smirnow, 1966, p.417):

$$f(t) = \sum_{k=0}^{\infty} (a_k \cos kt + b_k \sin kt) , \quad (31-1)$$

where a_k and b_k are coefficients; since $\sin kt = 0$ for $k = 0$, b_0 is arbitrary and will be put equal to zero.

This representation defines $f(t)$ also for arbitrary real t ($-\infty < t < \infty$) as a periodic function:

$$f(t \pm 2k\pi) = f(t) , \quad k = 1, 2, 3, \dots \quad (31-2)$$

In view of the well-known *orthogonality relations* of the trigonometric functions:

$$\begin{aligned} \int_0^{2\pi} \cos kt \cos lt dt &= 0 & \text{if } k \neq l , \\ \int_0^{2\pi} \sin kt \sin lt dt &= 0 & \text{if } k \neq l , \\ \int_0^{2\pi} \cos kt \sin lt dt &= 0 & \text{always ,} \\ \int_0^{2\pi} \cos^2 kt dt &= \int_0^{2\pi} \sin^2 kt dt = \pi & \text{if } k > 0 , \end{aligned} \quad (31-3)$$

the coefficients of the series (31-1) are given by

$$a_0 = \frac{1}{2\pi} \int_0^{2\pi} f(t) dt ,$$

$$\begin{aligned} a_k &= \frac{1}{\pi} \int_0^{2\pi} f(t) \cos kt \, dt & \text{if } k > 0, \\ b_k &= \frac{1}{\pi} \int_0^{2\pi} f(t) \sin kt \, dt & \text{if } k \geq 0. \end{aligned} \quad (31-4)$$

The function, defined in the xy -plane outside and on the unit circle,

$$f(x, y) = \sum_{k=0}^{\infty} r^{-k} (a_k \cos kt + b_k \sin kt) \quad (31-5)$$

with

$$r = \sqrt{x^2 + y^2}, \quad t = \arctan \frac{y}{x} \quad (31-6)$$

being polar coordinates, reduces on the unit circle $r = 1$ to (31-1) and is readily seen to be harmonic for $r > 1$, satisfying Laplace's equation

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = 0. \quad (31-7)$$

We thus have a very simple one-to-one relation between the function (31-1) defined on the unit circle and the harmonic function (31-5) defined outside; it will therefore be sufficient in the sequel to limit our study to (31-1).

A *stochastic process*, or *random function*, on the circle is a function $f(t, \omega)$ which depends, in addition to t , on a parameter ω which represents a "random choice". For any fixed value $\omega = \omega_1$ we get a function $f(t, \omega_1)$ of t only, which under the above-mentioned assumptions has the form (31-1); different ω_1 give different functions of t of form (31-1), which are considered as different "realizations" of the random process $f(t, \omega)$.

For instance, ω may denote the numbers 1, 2, 3, 4, 5, 6, so that $f(t, \omega)$ denotes 6 functions of form (31-1). By throwing a die we can determine ω (e.g., $\omega_1 = 5$) and the function $f(t, \omega_1)$ associated with it; this will explain the term, random function.

More generally, ω is a point in some *probability space*, or *sample space*, Ω . In this space we define a measure, such that measurable subsets of Ω are associated with events, the measure of a subset denoting the probability of the corresponding event. The measure of Ω itself is 1.

Let us illustrate this well-known fact, which can be found in any text-book on probability (the author's favorite is (Feller, 1957, 1966)) by means of the example just given, the throw of a die. Probability space Ω is the set of the six integers $\{1, 2, 3, 4, 5, 6\}$. Any of these integers, say 4, forms a subset of Ω , denoted by $\{4\}$. This subset corresponds to the event of throwing the face "4". To each of the subsets $\{1\}, \{2\}, \dots, \{6\}$ we associate the same measure $1/6$. The event of throwing a "2" or a "4" corresponds to the sum of the sets $\{2\}$ and $\{4\}$ and has probability equal to the sum of the individual probabilities:

$$\frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

The event of throwing a "1" or a "2" or a "3" or a "4" or a "5" or a "6" has the probability

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1,$$

that is, certainty, as it must be from an intuitive point of view: it is certain that one of the faces from "1" to "6" will show up. This illustrates the intuitive reason for demanding that the total measure of Ω is 1.

In this simple example we have 6 possible choices, or "sample points". In a more relevant case we need infinitely many possible choices, corresponding to a more general probability space Ω .

Let us return to our case of a random function on the circle

$$f = f(t, \omega), \quad 0 \leq t < 2\pi, \quad \omega \in \Omega, \quad (31-8)$$

Ω denoting a general probability space, which will be specialized later on. To get these simple but basic concepts firmly fixed in our mind, let us state again the meaning of the two arguments t and ω , using slightly different terms.

The variable t is the *space variable*, defining position in actual physical space. This becomes immediately evident on taking into account that the circle is a simplified analogue to the terrestrial sphere, so that a point on the circle, defined by t , corresponds to a point on the earth's surface.

On the other hand, ω , so to speak, describes chance: it defines a random choice. In statistical mechanics, the probability space Ω is called

phase space; we shall sometimes find this terminology convenient and call ω a *phase variable*. Anyway, ω serves as a kind of "random label" to distinguish one realization (or *sample function*) $f(t, \omega_1)$ of our stochastic process from another realization $f(t, \omega_2)$, both sample functions being functions of t only, since ω_1 or ω_2 are constants.

Generally speaking, a quantity depending on ω is called a *random variable*. This explains the name, random function, for a function $f(t, \omega)$ of t that depends, in addition, on "chance" ω .

Let us expand such a random function on the circle into a Fourier series (31-1) with respect to t . We have

$$f(t, \omega) = \sum_{k=0}^{\infty} [a_k(\omega) \cos kt + b_k(\omega) \sin kt] ; \quad (31-9)$$

clearly, the coefficients a_k and b_k will now be random variables depending on ω . By (31-4) they are given by

$$\begin{aligned} a_0(\omega) &= \frac{1}{2\pi} \int_0^{2\pi} f(t, \omega) dt , \\ a_k(\omega) &= \frac{1}{\pi} \int_0^{2\pi} f(t, \omega) \cos kt dt , \quad k > 0 , \end{aligned} \quad (31-10)$$

and similarly for $b_k(\omega)$.

32. THE COVARIANCE FUNCTION

Consider the values of a random function f at two different positions, t and $t + s$ (Fig. 32.1) and form their product:

$$f(t)f(t+s) ; \quad (32-1)$$

the dependence on ω will always be understood even if not explicitly written. A suitably defined average of the product (32-1) is nothing else than a *covariance function* corresponding to the random function $f(t) = f(t, \omega)$; it depends on the distance s and, possibly, also on t and ω .

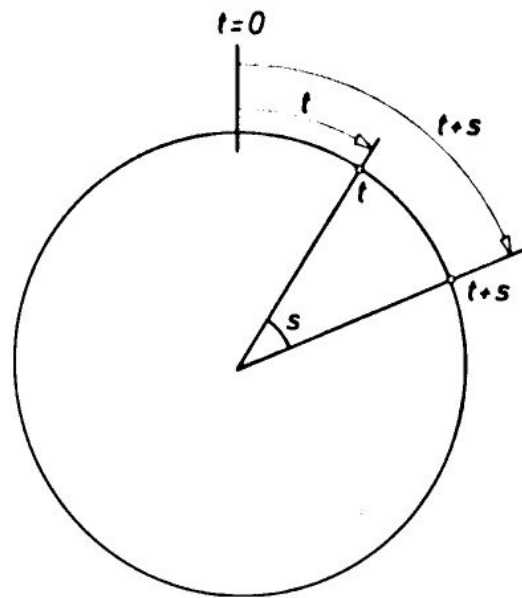


FIGURE 32.1. Positions on the circle.

For random functions, the natural definition of the average is in terms of the *statistical expectation* E :

$$\begin{aligned} C(s,t) &= E\{f(t)f(t+s)\} \\ &= \int_{\Omega} f(t,\omega)f(t+s,\omega) d\Omega ; \end{aligned} \quad (32-2)$$

E is defined as an integral over probability space Ω . This definition presupposes that the random function itself has zero expectation:

$$E\{f\} = \int_{\Omega} f(t,\omega) d\Omega = 0 . \quad (32-3)$$

By (31-10) this implies

$$E\{a_k\} = 0 = E\{b_k\} \quad \text{for all } k . \quad (32-4)$$

We substitute the Fourier expansion (31-9) into (32-2) and get

$$\begin{aligned}
C(s,t) &= E \left\{ \sum_{k=0}^{\infty} [a_k \cos kt + b_k \sin kt] \right. \\
&\quad \cdot \left. \sum_{l=0}^{\infty} [a_l \cos l(t+s) + b_l \sin l(t+s)] \right\} \\
&= E \left\{ \sum_{k,l} [a_k a_l \cos kt \cos l(t+s) + b_k b_l \sin kt \sin l(t+s) + \right. \\
&\quad \left. + a_k b_l \cos kt \sin l(t+s) + b_k a_l \sin kt \cos l(t+s)] \right\}. \quad (32-5)
\end{aligned}$$

The formal multiplication of the two Fourier series is justified since, by our assumption, they are uniformly convergent. For the same reason, we can perform the integration E term by term.

We shall now make the fundamental assumption that the Fourier coefficients are all *statistically uncorrelated*, that is, that all covariances between different coefficients vanish:

$$\begin{aligned}
E\{a_k a_l\} &= 0 \quad \text{if } k \neq l, \\
E\{b_k b_l\} &= 0 \quad \text{if } k \neq l, \\
E\{a_k b_l\} &= 0 \quad \text{always}.
\end{aligned} \quad (32-6)$$

We further assume that the variances of a_k and b_k , for each k , are equal:

$$E\{a_k^2\} = E\{b_k^2\} = c_k. \quad (32-7)$$

Then (32-5) becomes

$$\begin{aligned}
C(s,t) &= \sum_{k=0}^{\infty} [E\{a_k^2\} \cos kt \cos k(t+s) + \\
&\quad + E\{b_k^2\} \sin kt \sin k(t+s)].
\end{aligned} \quad (32-8)$$

In view of (32-7) and of the identity

$$\cos kt \cos k(t+s) + \sin kt \sin k(t+s) = \cos ks$$

this finally reduces to

$$C(s) = \sum_{k=0}^{\infty} c_k \cos ks, \quad (32-9)$$

which shows that the covariance function then depends on the distance s only. This function will be called the (*true*) *covariance function*.

The empirical covariance function. In practice, one frequently has only one realization of a stochastic process

$$f(t) = f(t, \omega), \quad \omega = \text{const.} \quad (32-10)$$

The question is whether it is possible to estimate the covariance function using this one sample function only.

In this case we cannot form the statistical expectation E , the *phase average* (if probability space Ω is denoted as phase space); instead, we form an average over t , the *space average* M (for stochastic processes on the real line, $-\infty < t < \infty$, t may be interpreted as time, so that M will be a "time average"). The space average M of (32-1) is defined as

$$r(s) = M\{f(t)f(t+s)\} = \frac{1}{2\pi} \int_0^{2\pi} f(t)f(t+s)dt, \quad (32-11)$$

$f(t)$ being, as always, understood as a periodic function (31-2). The function $r(s)$ is called the *empirical covariance function*.

In analogy to (32-3) we have the condition

$$M\{f(t)\} = \int_0^{2\pi} f(t)dt = 0, \quad (32-12)$$

which means by (31-4) that

$$a_0 = 0. \quad (32-13)$$

This is not an essential restriction since we can always replace $f(t)$ by $f(t) - a_0$, for which the zero-order coefficient is, in fact, zero. Thus we may assume (32-13) to hold.

Then the sum in (31-1) begins with $k = 1$, and substituting this series into (32-11) we get

$$\begin{aligned}
r(s) &= \frac{1}{2\pi} \int_0^{2\pi} \sum_{k=1}^{\infty} [a_k \cos kt + b_k \sin kt] \cdot \\
&\quad \cdot \sum_{l=1}^{\infty} [a_l \cos l(t+s) + b_l \sin l(t+s)] dt \\
&= \frac{1}{2\pi} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \int_0^{2\pi} [a_k a_l \cos kt \cos l(t+s) + \\
&\quad + b_k b_l \sin kt \sin l(t+s) + \\
&\quad + a_k b_l \cos kt \sin l(t+s) + \\
&\quad + b_k a_l \sin kt \cos l(t+s)] dt,
\end{aligned}$$

the formal operations (series multiplication and termwise integration) are again justified by uniform convergence.

The orthogonality relations (31-3) give at once:

$$\begin{aligned}
r(s) &= \frac{1}{2\pi} \sum_{k=1}^{\infty} \left[a_k^2 \int_0^{2\pi} \cos kt \cos k(t+s) dt + \right. \\
&\quad + b_k^2 \int_0^{2\pi} \sin kt \sin k(t+s) dt + \\
&\quad + a_k b_k \int_0^{2\pi} \cos kt \sin k(t+s) dt + \\
&\quad \left. + a_k b_k \int_0^{2\pi} \sin kt \cos k(t+s) dt \right], \tag{32-14}
\end{aligned}$$

since all products of trigonometric functions for $k \neq l$ vanish after integration.

We further have

$$\begin{aligned}
&\int_0^{2\pi} \cos kt \cos k(t+s) dt = \\
&= \int_0^{2\pi} (\cos^2 kt \cos ks - \cos kt \sin kt \sin ks) dt = \pi \cos ks,
\end{aligned}$$

again by (31-3), and similarly

$$\int_0^{2\pi} \sin kt \sin k(t+s) dt = \pi \cos ks.$$

Finally,

$$\begin{aligned} & \int_0^{2\pi} [\cos kt \sin k(t+s) + \sin kt \cos k(t+s)] dt = \\ & = \int_0^{2\pi} \sin k(2t+s) dt = 0 . \end{aligned}$$

Hence (32-14) reduces to

$$\Gamma(s) = \frac{1}{2} \sum_{k=1}^{\infty} (a_k^2 + b_k^2) \cos ks . \quad (32-15)$$

This is the Fourier expansion of the empirical covariance function. In view of (32-10), the function $f(t)$ and its Fourier coefficients a_k and b_k depend on ω :

$$a_k = a_k(\omega) , \quad b_k = b_k(\omega) . \quad (32-16)$$

Hence, also $\Gamma(s)$ depends on ω , so that (32-15) can be written more explicitly:

$$\Gamma(s, \omega) = \sum_{k=1}^{\infty} \gamma_k(\omega) \cos ks , \quad (32-17)$$

where

$$\gamma_k(\omega) = \frac{1}{2} [a_k^2(\omega) + b_k^2(\omega)] . \quad (32-18)$$

Let us now compare the empirical covariance function (32-15) or (32-17) with the true covariance function (32-9). Forming the expectation E of (32-18) we have

$$E\{\gamma_k\} = E\{\gamma_k(\omega)\} = \frac{1}{2} E\{a_k^2\} + \frac{1}{2} E\{b_k^2\} ,$$

so that by (32-7)

$$E\{\gamma_k\} = c_k . \quad (32-19)$$

The expectation of (32-17) is

$$E\{r(s, \omega)\} = \sum_{k=-1}^{\infty} E\{\gamma_k(\omega)\} \cos ks = \sum_{k=-1}^{\infty} c_k \cos ks ,$$

so that

$$E\{r(s)\} = C(s) , \quad (32-20)$$

the expectation of the empirical covariance function is the true covariance function. In statistical terms, the empirical covariance function is an unbiased estimate of the true covariance function.

It would be particularly desirable if the empirical covariance function is identical to the true covariance function, or if the two functions are equal at least for almost all ω (that is, for all ω with the possible exception of a set of measure zero). In this case, the covariance function can be exactly estimated from one realization of the stochastic process $f(t, \omega)$, that is, from one sample function $\omega = \text{const.}$ This is the case of *ergodicity*.

This name has been taken from statistical mechanics, where it means that a time average is the same as the corresponding phase average. In our case, the space average M of $f(t)f(t+s)$ should be equal to the phase average E of this product.

Obviously, ergodicity is a very special case, and the question arises whether it is possible at all. This question will be answered positively in the next section.

33. ERGODIC PROCESSES ON THE CIRCLE

The case in which the empirical covariance function $r(s, \omega)$ coincides, for almost all ω , with the true one, $C(s)$, has been called *ergodicity* in the preceding section. By comparing the coefficients of the respective Fourier expansions (32-9) and (32-15) we get the necessary and sufficient condition for ergodicity in this sense:

$$a_k^2(\omega) + b_k^2(\omega) = 2c_k$$

(33-1)

for almost all ω .

The meaning of this condition should be carefully kept in mind. The coefficients c_k , defined by (32-7), are given *nonrandom constants*. The coefficients a_k and b_k on the lefthand side are, however, functions of ω and hence *random variables*. Thus the condition (33-1) is certainly very restrictive.

It should be recalled that we have derived (33-1) under the assumption of uniform convergence of the Fourier series for $f(t, \omega)$. This assumption is not essential; for a proof under more general conditions (integrability) see (Zygmund, 1968, pp.36-37).

Lauritzen's theorem. In particular, it is impossible to satisfy the ergodicity condition by a stochastic process defined by (31-9) with $a_k(\omega)$ and $b_k(\omega)$ being uncorrelated and normally distributed (Gaussian) stochastic variables of zero expectation. This has been proved by Lauritzen (1973, p.65) by explicitly calculating the variance of the empirical covariance function $r(t, \omega)$ and showing that it is non-zero. (For ergodic processes this variance is evidently zero.)

For us, Lauritzen's theorem is an obvious, almost elementary consequence of (33-1). For Gaussian random variables, *uncorrelatedness* is equivalent to *statistical independence*. Hence (32-6) implies that all a_k and b_k are statistically independent random variables. If the functions $a_k(\omega)$ and $b_k(\omega)$ can vary independently of each other, then (33-1) can be violated at will. Eq. (33-1) would only be satisfied if

$$a_k(\omega) = \text{const.}, \quad b_k(\omega) = \text{const.} \quad (33-2)$$

for almost all ω , which is incompatible with zero expectation (32-4). These contradictions prove the theorem.

Lauritzen's theorem may be concisely, though somewhat loosely, formulated thus: *a Gaussian random process on the circle cannot be ergodic*. Looking for an ergodic process, we must, therefore, consider non-Gaussian processes. The a_k and b_k will be uncorrelated, but not necessarily statistically independent. It is known that statistical independence implies uncorrelatedness; the converse is true only for normal processes.

Ergodic processes: first example. Let the coefficients a_k and b_k , for different k , be statistically independent; for the same k , a_k and b_k will only be uncorrelated, in agreement with the third equation of (32-6). To satisfy the ergodicity condition (33-1), we take

$$a_k = \sqrt{2c_k} \cos \omega_k,$$

$$b_k = \sqrt{2c_k} \sin \omega_k, \quad (33-3)$$

where ω_k is a random variable uniformly distributed in the interval $0 \leq \omega_k < 2\pi$. This means that the probability density $\phi(\omega_k)$ of ω_k has the form of Fig. 33.1. Geometrically, a_k and b_k are represented in Fig. 33.2. We have a random vector with components $[a_k, b_k]$ of fixed length $\sqrt{2c_k}$ but with randomly variable azimuth ω_k . The end point of this vector thus describes a circle. Any point of the circle corresponds to a random choice of ω_k . The probability that the end point of the vector falls onto the arc AB is proportional to the arc length, namely $(\beta - \alpha)/2\pi$, corresponding to the shaded area in Fig. 33.1; this is the meaning of uniform distribution. Obviously, the probability that the end point lies somewhere on the circle is $2\pi/2\pi = 1$, namely certainty, as it should be.

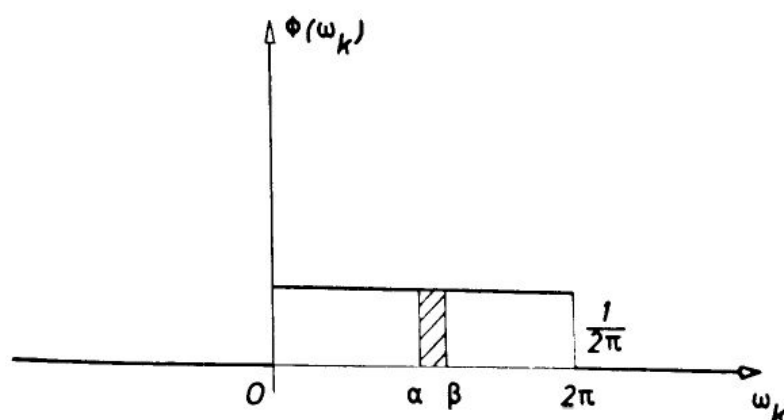


FIGURE 33.1. Rectangular distribution.

Thus the coefficients a_k and b_k are clearly statistically dependent, but are they still uncorrelated? We have

$$E\{a_k b_k\} = \int_0^{2\pi} a_k(\omega_k) b_k(\omega_k) \phi_k(\omega_k) d\omega_k \quad (33-4)$$

where

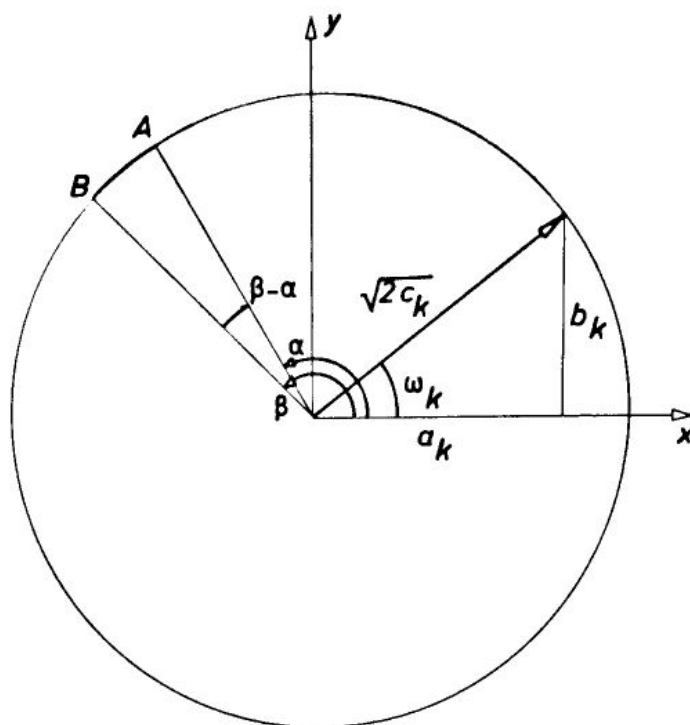


FIGURE 33.2. Random azimuth.

$$\phi_k(\omega) = \frac{1}{2\pi}, \quad 0 \leq \omega_k < 2\pi, \quad (33-5)$$

so that by (33-3)

$$\begin{aligned} E\{a_k b_k\} &= \int_0^{2\pi} 2c_k \cos \omega_k \sin \omega_k \frac{1}{2\pi} d\omega_k \\ &= \frac{c_k}{2\pi} \int_0^{2\pi} \sin 2\omega_k d\omega_k = 0; \end{aligned} \quad (33-6)$$

hence a_k and b_k are, in fact, uncorrelated. This provides a simple geometrical illustration of how two random variables can be uncorrelated without being statistically independent.

For different k , the vectors $[a_k, b_k]$ have been supposed to be independent. This means that the ω_k are uniformly distributed, independent random variables. Consider two ω_k , say, ω_2 and ω_3 . Each ω_k varies from 0 to 2π , that is, over the (unit) circle. Since the joint probability space of two independent random variables is the cartesian product of the two individual probability spaces, the joint probability space of

ω_1 and ω_2 is the cartesian product of two circles. The joint probability space of $[\omega_1, \omega_2, \dots, \omega_n]$ is the product of n circles.

The Fourier series of the random function $f(t, \omega)$ involves infinitely many coefficients a_k and b_k . The probabilistic event of sorting out one sample function thus requires infinitely many independent choices of $\omega_1, \omega_2, \omega_3, \dots$. The probability space for $f(t, \omega)$ is, therefore, the cartesian product of infinitely many circles, or ω represents the infinite vector

$$\omega = [\omega_1, \omega_2, \omega_3, \dots], \quad (33-7)$$

each ω_k being independently uniformly distributed.

Finally we prove that if one sample function of our present ergodic process has a uniformly convergent Fourier series, then this will hold for all sample functions of this process. Since the absolute values of sine and cosine are, at most, equal to 1, the Fourier series (31-1), with $a_0 = 0$, has the majorant

$$\sum_{k=1}^{\infty} (|a_k| + |b_k|). \quad (33-8)$$

Convergence of this majorant series is clearly sufficient for uniform convergence of our Fourier series; that it is also necessary is a consequence of the Theorem of Denjoy-Lusin (Zygmund, 1968, p.232).

Since

$$\sqrt{a^2 + b^2} \leq |a| + |b| \leq 2\sqrt{a^2 + b^2}, \quad (33-9)$$

convergence of (33-8) is logically equivalent to the convergence of

$$\sum_{k=1}^{\infty} \sqrt{a_k^2 + b_k^2} = \sum_{k=1}^{\infty} \sqrt{2c_k}, \quad (33-10)$$

which is, therefore, also a necessary and sufficient condition for the uniform convergence of our Fourier series. Therefore, uniform convergence of the Fourier series of *one* sample function implies convergence of the right-hand side of (33-10). Since this right-hand side does not depend on ω , the left-hand side of this equation must converge for all ω , which implies uniform convergence of the Fourier series of the sample functions for any ω , which was to be shown.

The uniform distribution on a circle is even simpler than a normal distribution. Furthermore, the "probability circle" $0 \leq \omega_k < 2\pi$ seems, somehow, to be a natural counterpart of the "space circle" $0 \leq t < 2\pi$. Thus the present simple example seems to be a quite natural model for a stochastic process on the circle, more natural than any Gaussian model; furthermore it is ergodic. The next example is still simpler.

Ergodic processes: second example. We now take ω itself as a random variable uniformly distributed in the interval

$$0 \leq \omega < 2\pi, \quad (33-11)$$

or, what is the same, on the unit circle. Thus, in the random function $f(t, \omega)$, both variables t and ω now range over a unit circle, the circle for t representing "ordinary space" and the circle for ω representing "probability space".

We now take

$$f(t, \omega) = f(t + \omega). \quad (33-12)$$

Let

$$f(t, 0) = f(t) \quad (33-13)$$

be one realization of the stochastic process, for $\omega = 0$; we shall call it the *initial realization*. Any other realization $f(t, \omega) = f(t + \omega)$ represents simply a rotation of the circle, or of the function $f(t)$, by the angle ω (Fig.33.3).

We may also write

$$f(t + \omega) = R_\omega f(t), \quad (33-14)$$

where the operator R_ω means rotation by the angle ω . In other terms, we may identify our probability space (33-11) with rotation group space. In fact, in the plane, the rotation group is one-dimensional, being characterized by one angle ω .

The functions $f(t, \omega)$ differ from each other only by a rotation; they are not essentially different (Fig.33.3). This model, therefore, is suited to represent the case in which there is only one realization $f(t)$ and we wish to use the mathematical techniques of stochastic processes; this is

Justified in the case of homogeneity, if a rotation by ω gives a physically equally meaningful situation, so that $f(t+\omega)$ instead of $f(t)$ would be physically equally possible.

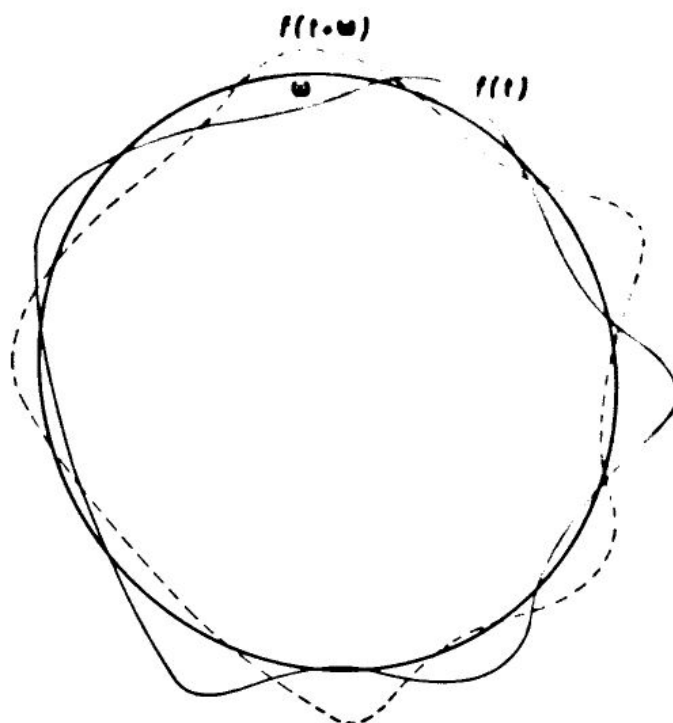


FIGURE 33.3. The rotation $f(t) \rightarrow f(t+\omega)$.

We assume (32-12) to hold for our initial $f(t)$. Then (32-3) becomes

$$\begin{aligned} E\{f(t, \omega)\} &= \frac{1}{2\pi} \int_0^{2\pi} f(t, \omega) d\omega \\ &= \frac{1}{2\pi} \int_0^{2\pi} f(t+\omega) d\omega. \end{aligned} \quad (33-15)$$

We change the integration variable by putting (t is constant with respect to integration!)

$$t + \omega = u, \quad d\omega = du, \quad (33-16)$$

obtaining

$$\begin{aligned}
 E(f(t, \omega)) &= \frac{1}{2\pi} \int_t^{t+2\pi} f(u) du \\
 &= \frac{1}{2\pi} \int_0^{2\pi} f(u) du = 0
 \end{aligned}
 \tag{33-17}$$

by (32-12), so that (32-3) is satisfied.

Now the covariance function (32-2) becomes

$$C(s, t) = \frac{1}{2\pi} \int_0^{2\pi} f(t+\omega) f(t+s+\omega) d\omega . \tag{33-18}$$

The substitution (33-16) transforms this integral into

$$\begin{aligned}
 &\frac{1}{2\pi} \int_0^{2\pi} f(u) f(u+s) du \\
 &= \frac{1}{2\pi} \int_0^{2\pi} f(t) f(t+s) dt = r(s)
 \end{aligned}
 \tag{33-19}$$

by (32-11). Therefore, in this model, the *empirical covariance function coincides with the true covariance function*; the process is ergodic. In fact, the "phase average" E is seen to coincide with the "space average" M . Since E can be transformed into M by a simple change of variables, the process under consideration is *trivially ergodic*.

This is obviously a very simple situation, but an important one. We shall, therefore, try to understand it better by studying the spectral representation, that is, the Fourier series expansion.

We shall denote the Fourier coefficients of the initial representation $f(t) = f(t, 0)$ by a_k and b_k . The coefficients $a_k(\omega)$ and $b_k(\omega)$ of $f(t, \omega)$ are then given by (31-10). We have

$$a_0(\omega) = \frac{1}{2\pi} \int_0^{2\pi} f(t+\omega) dt = \frac{1}{2\pi} \int_0^{2\pi} f(t) dt = 0 \tag{33-20}$$

by (32-12). For $k > 0$ we get

$$a_k(\omega) = \frac{1}{\pi} \int_0^{2\pi} f(t+\omega) \cos kt dt \tag{33-21}$$

and substituting (ω is constant with respect to integration)

$$t + \omega = v, \quad dt = dv, \quad (33-22)$$

we have

$$\begin{aligned} a_k(\omega) &= \frac{1}{\pi} \int_0^{2\pi} f(v) \cos k(v-\omega) dv \\ &= \frac{1}{\pi} \int_0^{2\pi} f(v) (\cos kv \cos k\omega + \sin kv \sin k\omega) dv \\ &= \cos k\omega \cdot \frac{1}{\pi} \int_0^{2\pi} f(v) \cos kv dv + \\ &\quad + \sin k\omega \cdot \frac{1}{\pi} \int_0^{2\pi} f(v) \sin kv dv \end{aligned} \quad (33-23)$$

or, by (31-4),

$$a_k(\omega) = a_k \cos k\omega + b_k \sin k\omega. \quad (33-24)$$

In exactly the same way, replacing $\cos kt$ by $\sin kt = \sin k(v-\omega)$ we get

$$b_k(\omega) = -a_k \sin k\omega + b_k \cos k\omega. \quad (33-25)$$

Let us now evaluate (32-6), using (33-24) and (33-25). We get

$$\begin{aligned} E\{a_k(\omega)a_1(\omega)\} &= \frac{1}{2\pi} \int_0^{2\pi} a_k(\omega)a_1(\omega)d\omega \\ &= \frac{1}{2\pi} \int_0^{2\pi} (a_k \cos k\omega + b_k \sin k\omega) \cdot \\ &\quad \cdot (a_1 \cos 1\omega + b_1 \sin 1\omega) d\omega. \end{aligned}$$

On termwise multiplication and integration, using the orthogonality relations (31-3), we readily obtain the value zero if $k \neq 1$. Proceeding similarly, we see that all orthogonality relations (32-6) are satisfied.

We further obtain

$$\begin{aligned}
 E(a_k(\omega)^2) &= \frac{1}{2\pi} \int_0^{2\pi} (a_k \cos k\omega + b_k \sin k\omega)^2 d\omega \\
 &= \frac{1}{2}(a_k^2 + b_k^2) .
 \end{aligned}$$

(33-24)

In the same way,

$$E(b_k^2(\omega)) = \frac{1}{2}(a_k^2 + b_k^2) ,$$

(33-25)

which is independent of ω , so that (32-7) is satisfied with

$$c_k = \frac{1}{2}(a_k^2 + b_k^2) .$$

(33-26)

We finally compute, using (33-24) and (33-25),

$$\begin{aligned}
 a_k(\omega)^2 + b_k(\omega)^2 &= (a_k \cos k\omega + b_k \sin k\omega)^2 + \\
 &\quad + (-a_k \sin k\omega + b_k \cos k\omega)^2 \\
 &= a_k^2 + b_k^2 .
 \end{aligned}$$

Thus

$$a_k(\omega)^2 + b_k(\omega)^2 = 2c_k ,$$

which shows that the ergodicity condition (33-1) is, in fact, satisfied.

Let us finally compare this model with our first ergodic model. In the first model, probability space is the cartesian product of infinitely many circles $0 \leq \omega_k < 2\pi$ ($k = 1, 2, 3, \dots$), ω being the infinite vector (33-7), consisting of independent, uniformly distributed random variables. In the present model, probability space is simply one circle $0 \leq \omega < 2\pi$, ω being a uniformly distributed one-dimensional random variable. Therefore, in the first model, a_k and a_l for $k \neq l$, depending on different independent random variables ω_k and ω_l , are statistically independent. The same holds for a_k and b_l , and for b_k and b_l . For the same k , a_k and b_k are dependent though uncorrelated. On the other hand, in the present model, all Fourier coefficients depend on the same variable ω ;

therefore, all are statistically dependent, but, as we have seen, any two different coefficients are uncorrelated, as a consequence of the orthogonality relations (31-3) for trigonometric functions.

34. STOCHASTIC PROCESSES ON THE SPHERE

Notations. Our preceding considerations about stochastic processes on the circle can be translated almost literally to the sphere. Instead of the Fourier series (31-1) we have the spherical-harmonic series

$$f(\theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=0}^n [a_{nm} R_{nm}(\theta, \lambda) + b_{nm} S_{nm}(\theta, \lambda)] \quad (34-1)$$

where

$$\begin{aligned} R_{nm}(\theta, \lambda) &= P_{nm}(\cos \theta) \cos m\lambda, \\ S_{nm}(\theta, \lambda) &= P_{nm}(\cos \theta) \sin m\lambda \end{aligned} \quad (34-2)$$

as usual (sec.3). The function $f(\theta, \lambda)$ may be interpreted as the anomalous potential T or the gravity anomaly Δg .

To simplify the notation, let us put

$$S_{nm}(\theta, \lambda) = R_{n,-m}(\theta, \lambda), \quad m = 1, 2, \dots, n, \quad (34-3)$$

so that any R_{nm} with negative second subscript denotes the corresponding S_{nm} , for instance, $R_{5,-3} = S_{53}$. Then (34-1) may be simply written as

$$f(\theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=-n}^n a_{nm} R_{nm}(\theta, \lambda), \quad (34-4)$$

if for the coefficients we use an analogous notational convention:

$$a_{n,-m} = b_{nm}, \quad m = 1, 2, \dots, n. \quad (34-5)$$

It will be convenient also to use fully normalized harmonics, denoted by \bar{R}_{nm} and \bar{S}_{nm} , or by \bar{R}_{nm} with $-n \leq m \leq n$, which differ from the conventional harmonics by a factor and are normalized by

$$\frac{1}{4\pi} \iint_{\sigma} \bar{R}_{nm}^2 d\sigma = 1, \quad (34-6)$$

σ denoting the unit sphere; cf. eq. (3-27). The orthonormality relations may be written

$$\overline{M}(\bar{R}_{nm} \bar{R}_{qp}) = \delta_{nq} \delta_{mp}, \quad (34-7)$$

where

$$\overline{M}\{\cdot\} = \frac{1}{4\pi} \iint_{\sigma} (\cdot) d\sigma \quad (34-8)$$

denotes now the average over the unit sphere and δ_{kl} is the Kronecker delta, 1 if $k = l$ and 0 otherwise.

If we write (34-4) in fully normalized harmonics,

$$f(\theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \bar{a}_{nm} \bar{R}_{nm}(\theta, \lambda), \quad (34-9)$$

then the coefficients are simply given by

$$\bar{a}_{nm} = \overline{M}\{f \bar{R}_{nm}\}, \quad (34-10)$$

in view of the orthonormality; these equations are a shorthand notation of eqs. (3-29).

As a final notational convention regarding spherical harmonic expansions, we introduce the two-dimensional parameter

$$t = [\theta, \lambda] \quad (34-11)$$

and write (34-9) as

$$f(t) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \bar{a}_{nm} \bar{R}_{nm}(t); \quad (34-12)$$

this stresses the analogy with the case of the circle.

The stochastic parameter will again be denoted by $\omega \in \Omega$, Ω being probability space with total measure 1. Then

$$f(t, \omega)$$

will denote a stochastic process on the sphere. The expectation E is again defined by

$$E(\cdot) = \int_{\Omega} (\cdot) d\Omega \quad (34-13)$$

as an average over probability space, or *phase average*, as opposed to the *space average* \bar{M} defined by (34-8).

In analogy to (31-5), there is a one-to-one correspondence between continuous functions on the sphere and harmonic functions in space: the spatial function

$$f(r, \theta, \lambda) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{a_{nm}}{r^{n+1}} R_{nm}(\theta, \lambda) \quad (34-14)$$

satisfies Laplace's equation outside σ . Therefore, there is a one-to-one correspondence between harmonic stochastic processes in space and stochastic processes on the sphere, and we can limit our considerations to the latter. This sphere may be identified with sea level (p.98).

Covariances. We again assume that our stochastic process is centered:

$$E\{f(t, \omega)\} = 0. \quad (34-15)$$

Then the covariance $C(t, u)$ between $f(t, \omega)$ and $f(u, \omega)$ at two different points t and u on the unit sphere σ is, as usual, defined by

$$C(t, u) = E\{f(t)f(u)\}, \quad (34-16)$$

the dependence on ω being understood.

As in the circular case, we shall limit ourselves to continuously differentiable functions. Then the spherical-harmonic expansion will be a uniformly convergent series (Kellogg, 1929, p.259), which can be multiplied and termwise integrated.

We, therefore, substitute (34-12) into (34-16):

$$C(t, u) = E \left\{ \sum_{n=0}^{\infty} \sum_{m=-n}^n \bar{a}_{nm} \bar{R}_{nm}(t) \cdot \sum_{q=0}^{\infty} \sum_{p=-q}^q \bar{a}_{qp} \bar{R}_{qp}(u) \right\},$$

multiply and integrate termwise with respect to ω (that is, interchange the order of summation and integration), obtaining

$$C(t,u) = \sum_n \sum_m \sum_q \sum_p E(\bar{a}_{nm} \bar{a}_{qp}) \bar{R}_{nm}(t) \bar{R}_{qp}(u) . \quad (34-17)$$

Let us now assume that the coefficients $\bar{a}_{nm} = \bar{a}_{nm}(\omega)$ are mutually uncorrelated random variables:

$$E(\bar{a}_{nm} \bar{a}_{qp}) = 0 \quad (34-18)$$

if $q \neq n$ or $p \neq m$ or both, and that $E(\bar{a}_{nm}^2)$ is the same for all coefficients of degree n , that is, for all m ; we put

$$E\{\bar{a}_{nm}^2\} = \frac{c_n}{2n+1} . \quad (34-19)$$

Then (34-17) becomes

$$C(t,u) = \sum_{n=0}^{\infty} \sum_{m=-n}^n \frac{c_n}{2n+1} \bar{R}_{nm}(t) \bar{R}_{nm}(u) . \quad (34-20)$$

Now we make use of the decomposition formula (3-30), which in our present notation takes the form

$$P_n(\cos\psi) = \frac{1}{2n+1} \sum_{m=-n}^n \bar{R}_{nm}(t) \bar{R}_{nm}(u) \quad (34-21)$$

with

$$t = [\theta, \lambda] \quad \text{and} \quad u = [\theta', \lambda'] , \quad (34-22)$$

ψ being the spherical distance between the points t and u :

$$\cos\psi = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\lambda'-\lambda) , \quad (34-23)$$

and $P_n(\cos\psi)$ denoting the (conventional) Legendre polynomial of degree n . Thus (34-20) reduces to

$$C(\psi) = \sum_{n=0}^{\infty} c_n P_n(\cos \psi) \quad (34-24)$$

Thus, the covariance function depends only on the spherical distance ψ . This is the important case of *homogeneity and isotropy*; it is seen to result from the postulate that the variances (34-19) of all coefficients \bar{a}_{nm} of the same degree n are equal.

The empirical covariance function. If there is only one realization of the stochastic process, we cannot directly compute the true covariance function C defined by (34-16) and expressed by (34-24). We may again try to compute an empirical covariance function r by replacing the phase average E by a suitable space average and hope that r will be a good estimate of C ; if possible, r should even be equal to C .

In view of the homogeneity and isotropy, we must integrate not only over the sphere (homogeneity), but in addition over the azimuth (isotropy). Therefore, we must supplement the average \bar{M} , defined by (34-8), by additionally averaging over the azimuth α . The resulting average M may be defined by

$$M\{\cdot\} = \frac{1}{8\pi^2} \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} \int_{\alpha=0}^{2\pi} (\cdot) \sin \theta d\theta d\lambda d\alpha \quad (34-25)$$

The geometric situation is shown by Fig. 34.1. The averaging is first performed over the circle $\psi = \text{const.}$, whose center $t = (\theta, \lambda)$ is then made to vary over the whole sphere σ .

This definition of M has already been used before; cf. eq. (10-2).

The angles λ, θ, α can be regarded as the three Eulerian angles defining a rotation in three-dimensional space, that is, λ, θ, α are the coordinates of an element of the rotation group or, of a "point" in "rotation group space". Therefore, M will be called a *rotation group average*.

Hence the empirical covariance function is given by

$$r(\psi) = M\{f(t)f(u)\} \quad (34-26)$$

where M is defined by (34-25) and the points t and u have the spherical distance ψ , which is constant with respect to the integration. If $f(\theta, \lambda)$ denotes the anomalous potential T , then $r(\psi)$ coincides with the function $K(\psi)$ as given by (10-2).

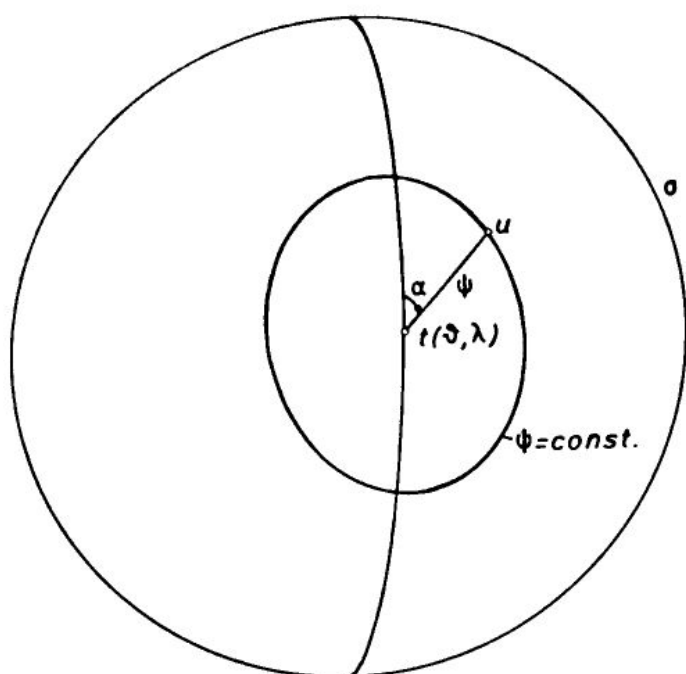


FIGURE 34.1. Integration over rotation group space.

Because of the way in which the average M is computed, the empirical covariance function will depend only on the distance ψ and can, therefore, be expanded into a series of Legendre polynomials of ψ :

$$\Gamma(\psi) = \sum_{n=0}^{\infty} \gamma_n P_n(\cos \psi) . \quad (34-27)$$

The γ_n can be expressed in terms of the spherical-harmonic coefficients \bar{a}_{nm} of the same n , in full analogy to (32-18). This is accomplished by eq. (10-8), which in the present notation reads

$$\gamma_n = \sum_{m=-n}^n \bar{a}_{nm}^2 . \quad (34-28)$$

Note again that this very simple expression is obtained by using conventional harmonics on the left-hand side and fully normalized harmonics on the right-hand side.

Clearly, γ_n , as well as \bar{a}_{nm} , are random variables, that is, functions of ω . Their expectation is given by

$$E\{\gamma_n\} = E\{\gamma_n(\omega)\} = \sum_{m=-n}^n E\{\bar{a}_{nm}^2(\omega)\}.$$

In view of (34-19) this becomes

$$E\{\gamma_n\} = c_n, \quad (34-29)$$

so that

$$E\{\Gamma(\psi)\} = C(\psi), \quad (34-30)$$

exactly as in the circular case (32-20): $\Gamma(\psi)$ is an unbiased estimate of $C(\psi)$.

35. ERGODIC PROCESSES ON THE SPHERE

For an ergodic process, $\Gamma(\psi)$ coincides with $C(\psi)$. The ergodicity condition, corresponding to (33-1), is

$$\sum_{m=-n}^n \bar{a}_{nm}^2(\omega) = c_n, \quad (35-1)$$

for almost all ω , c_n is independent of ω . This condition is equivalent to $\gamma_n = c_n$.

Lauritzen's theorem. Assume that $\bar{a}_{nm}(\omega)$ are normally distributed (Gaussian) random variables. For Gaussian variables, uncorrelatedness is equivalent to statistical independence. From our basic presupposition (34-18) it thus follows that the $\bar{a}_{nm}(\omega)$ must be statistically independent of each other. Then the summands on the left-hand side are independent functions of ω , so that (35-1) will be violated for almost all ω . Loosely formulated: a Gaussian random process on the sphere cannot be ergodic.

The present simple proof of Lauritzen's theorem suffers from the slight logical defect that (35-1) has been derived on the assumption that our stochastic process is sufficiently smooth (differentiable). Since Gaussian random variables may take arbitrarily large values, the convergence of the corresponding spherical-harmonic series cannot be guaranteed; still less are we sure that the corresponding realizations will all be differentiable (this may even be a practical argument against admitting a Gaussian process

as a mathematical model for the terrestrial gravity field). In fact, (35-1), just as (33-1), holds for more general assumptions, but we have not proved this because, for the present ergodic models, differentiability can be presupposed.

Thus, our deduction of Lauritzen's theorem has the character of a heuristic argument rather than of a fully rigorous mathematical proof, which is given in (Lauritzen, 1973, p.65). It has, however, the decisive advantage of showing the essential statistical situation underlying it, and the fact that the Gaussian character of the process is essential to the theorem: only for Gaussian distributions does uncorrelatedness imply statistical independence.

We shall now consider two examples of (non-Gaussian) ergodic stochastic processes on the sphere, corresponding to the two examples for the circle given in sec. 33.

First example: uniformly distributed coefficients. The two Fourier coefficients a_k and b_k define a two-dimensional vector whose end point lies on a circle of radius $\sqrt{2c_k}$ (Fig.33.2). Similarly, the $2n+1$ coefficients \bar{a}_{nm} (n fixed, $-n \leq m \leq n$) form a $(2n+1)$ -dimensional vector

$$\underline{a} = [\bar{a}_{n,-n}, \bar{a}_{n,-n+1}, \dots, \bar{a}_{n,n-1}, \bar{a}_{n,n}] \quad (35-2)$$

whose end point lies on a sphere of radius $\sqrt{c_n}$ in R^{2n+1} (Euclidean space of dimension $2n+1$); in fact, (35-1) may be written

$$|\underline{a}|^2 = c_n. \quad (35-3)$$

Assume now that different realizations of the stochastic process correspond to different positions of the endpoint of \underline{a} on this sphere. In other terms, if \underline{e} is the unit vector corresponding to \underline{a} , then

$$\underline{a}(\omega) = \sqrt{c_n} \underline{e}(\omega), \quad (35-4)$$

the unit vector being a function of ω : the random vector \underline{a} has a random direction but a constant length, in complete analogy to (33-3). The random directions $\underline{e}(\omega)$ are *uniformly distributed* in our $(2n+1)$ -dimensional space: probability is given by an area on the unit sphere in this space; cf. (Feller, 1966, p.68) for R^3 .

In view of (35-3), the $2n+1$ coefficients \bar{a}_{nm} of the same degree n are statistically dependent, but they are uncorrelated: it is easy to see

that (34-18) holds for them. In fact, this means that, for two different components of the vector \underline{a} or of the vector \underline{e} , say e_i and e_j , the integral, over the unit sphere σ_{2n} in R^{2n+1} , of its product $e_i e_j$ is zero:

$$\int_{\sigma_{2n}} e_i e_j d\sigma_{2n} = 0. \quad (35-5)$$

Denote the left-hand side of this equation by Q_{ij} :

$$Q_{ij} = \int_{\sigma_{2n}} e_i e_j d\sigma_{2n}; \quad (35-6)$$

we must prove that Q_{ij} is zero.

In fact, it follows from the definition (35-6) that Q_{ij} is invariant with respect to an interchange of the two axes x_i and x_j ; it is thus the same regardless of whether the coordinate system is right-handed or left-handed. Because of the spherical symmetry, Q_{ij} is invariant with respect to rotation and to reflection; it only depends on the geometrical configuration. This geometrical configuration-- $2n+1$ mutually orthogonal axes--remains unchanged if we replace the x_j -axis by its opposite direction giving

$$\int_{\sigma_{2n}} e_i (-e_j) d\sigma_{2n} = -Q_{ij},$$

which must, therefore, be equal to Q_{ij} . From $Q_{ij} = -Q_{ij}$ we get $Q_{ij} = 0$ and hence (35-5).

The reader is invited to make this reasoning clear to himself for the case of three-dimensional space with $i = 1$ and $j = 2$. (In this case, (35-5) is equivalent to the orthogonality of the first-degree harmonics. Why?)

So far, we have restricted our considerations to the $2n+1$ coefficients a_{nm} corresponding to the same degree n . Let us now consider two different degrees, say n and n' . Any two coefficients a_{nm} belonging to two different degrees will be assumed to be stochastically independent. Thus the probability space Ω is the cartesian product of infinitely many unit spheres:

$$\Omega = \sigma_2 \times \sigma_4 \times \sigma_6 \times \sigma_8 \times \dots \times \sigma_{2n} \times \sigma_{2n+2} \times \dots \quad (35-7)$$

where σ_{2n} denotes the $2n$ -dimensional unit sphere in R^{2n+1} . Thus the dimensionality of the spheres increases with increasing n , in contrast to the circular case where the probability space is the cartesian product of infinitely many identical circles.

To repeat: in the present model any two different a_{nm} are uncorrelated, but for different reasons: if the two coefficients belong to different degrees, then they are uncorrelated as a consequence of the statistical independence; if the two coefficients belong to the same degree n , then they are uncorrelated because of the orthogonality relation (35-5).

A second model of an ergodic stochastic process is obtained by taking the probability space Ω as rotation group space; this is the three-dimensional analogue of the second example of an ergodic process considered in sec. 33. In view of its basic importance we shall devote the next section to this model.

36. ROTATION GROUP SPACE

As we have seen in sec. 33, rotations of the circle, which constitute the rotation group in two dimensions, are described by one parameter ω ranging from 0 to 2π : the group of rotations of the plane forms a one-dimensional space, which may be identified with the unit circle.

Rotations of the sphere, which make up the rotation group in three dimensions, are described by three parameters, for which we may take three Eulerian angles: the group of rotations of three-dimensional space forms itself a three-dimensional space, whose coordinates are the three Eulerian angles. This three-dimensional rotation group space cannot be identified with the unit sphere. This is in contrast to the case of rotations of the circle and accounts for the greater complexity of the present case.

Various authors use various definitions of Eulerian angles. The following definition is fairly widely used and is best suited for the present purpose because of its relation to the spherical coordinates θ, λ .

Let a rectangular coordinate system XYZ be rotated into a position xyz by a general spatial rotation. This rotation is split up into three successive rotations around coordinate axes. The first rotation is about the Z -axis through an angle Λ ; it transforms the XYZ -system into X_1Y_1Z . The second rotation is about the Y_1 -axis through an angle θ ; thus we obtain a system $X_2Y_1Z_2$. Finally we rotate about the Z_2 -axis about an angle ψ in a positive, or $-\psi$ in the negative, sense, to obtain the desired system xyz .

The three angles Λ, θ, ψ are the Euler angles. They may be illustrated

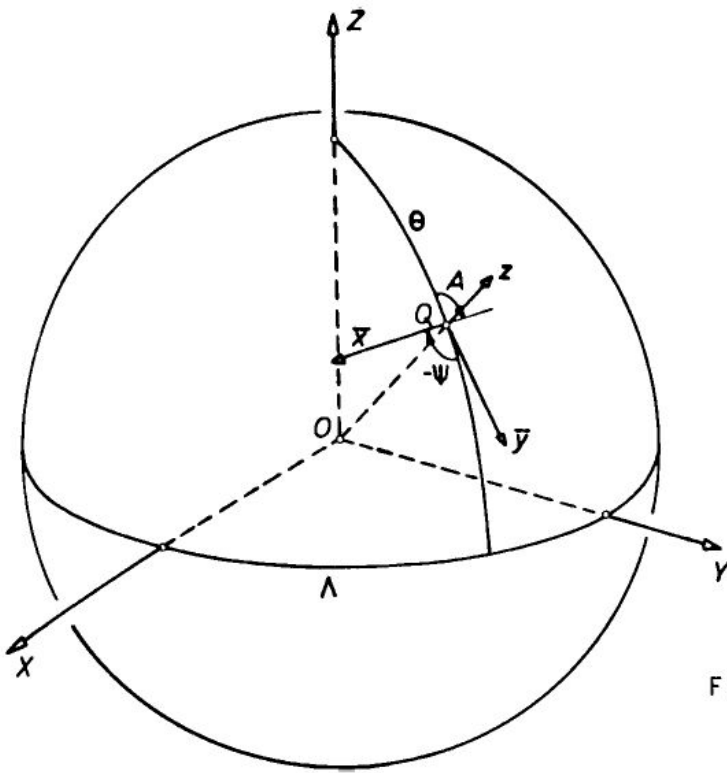


FIGURE 36.1. The Euler angles λ, θ, ψ .

in the following way (Fig.36.1). The angles θ and λ are the usual polar coordinates of the new z -axis. Let Q be the point in which the z -axis intersects the unit sphere, and denote by \bar{x} and \bar{y} the parallels through Q to the x - and y -axis, respectively. Then ψ is the angle which \bar{x} forms with the meridian, positive when counted counterclockwise (the figure shows a negative ψ).

In the usual terminology, the angle $-\psi$ is nothing else than the azimuth, counted clockwise, of the negative \bar{x} -direction. We shall, therefore, put

$$A = -\psi \quad (36-1)$$

and consider θ, λ, A as our final Eulerian angles.

These three angles define a point ω in rotation group space, which we shall denote by Ω (we shall later interpret it as our probability space); we thus put

$$\omega = [\theta, \lambda, A] . \quad (36-2)$$

The respective ranges are

$$0 \leq \theta \leq \pi , \quad 0 \leq \lambda < 2\pi , \quad 0 \leq A < 2\pi . \quad (36-3)$$

A rotation defined by the three Euler angles (36-2) will be denoted by R_ω . The value

$$\omega_0 = [0, 0, 0] \quad (36-4)$$

corresponds to the identity transformation I , leaving the axes XYZ unchanged; symbolically,

$$R_0 = I. \quad (36-5)$$

A unit vector t defined by spherical coordinates θ, λ has the components

$$t = \begin{bmatrix} \sin \theta \cos \lambda \\ \sin \theta \sin \lambda \\ \cos \theta \end{bmatrix}. \quad (36-6)$$

It will be symbolically abbreviated as

$$t = [\theta, \lambda]; \quad (36-7)$$

this notation has already been used before; cf. equation (34-11).

The rotation (36-2) transforms the vector t into another unit vector, which we shall denote by

$$R_\omega t; \quad (36-8)$$

it is convenient to consider R_ω as a rotation matrix, so that (36-8) is the usual product of a matrix and a vector.

Clearly, the Euler angles θ, λ or the rotation R_ω are completely different from and independent of the coordinates θ, λ of the vector t . There is, however, an interesting relation between these two sets of quantities. Form the triple

$$\tau = [\theta, \lambda, \alpha], \quad (36-9)$$

with an arbitrary value α between 0 and 2π , and

$$-\tau = [-\theta, -\lambda, -\alpha]. \quad (36-10)$$

Then it is easily seen that

$$R_{-t} t = e_z, \quad (36-11)$$

which is the unit vector of the Z-axis. Thus, the operation R_{-t} rotates an arbitrary unit vector t into the Z-axis. This simple fact will be of importance later on.

After these introductory geometrical considerations we are in a position to construct our stochastic process. We take a basic function

$$f(t) = f(\theta, \lambda) \quad (36-12)$$

on the unit sphere and define our stochastic process by

$$f(t, \omega) = f(R_\omega t). \quad (36-13)$$

This is in analogy to the two-dimensional case, equations (33-12) and (33-14); we shall also follow the respective developments in sec. 33 as closely as possible.

Again, the functions $f(t, \omega)$ differ from each other only by a rotation of the sphere; they are not essentially different. Our model is suited to represent the case in which we have only one realization $f(t)$ but wish to formally use the mathematical techniques of stochastic processes. This is the case of the terrestrial gravitational field, if we take

$$f(t) = T(\theta, \lambda), \quad (36-14)$$

which is the anomalous potential at sea level. The choice (36-13) is intimately connected with homogeneity and isotropy, i.e., with invariance of essential features with respect to rotations R_ω . More about this will be said in sec. 38.

According to (36-13), probability space Ω is rotation group space, a point $\omega \in \Omega$ being defined by the three Eulerian angles (36-2). The expectation

$$E\{\cdot\} = \iiint_{\Omega} (\cdot) d\Omega \quad (36-15)$$

is an integral over rotation group space. The integration is to be extended over the range (36-3); the problem is to find a suitable volume element $d\Omega$, defining a probability measure.

The product of two rotations R is again a rotation. The vector

$$R_1 R_2 t \quad (36-16)$$

is obtained by rotating the vector first by the matrix R_2 and then by the matrix R_1 . Assume that a spherical triangle $P_1 P_2 P_3$ is brought by a rotation R_1 into the position $P'_1 P'_2 P'_3$; the configuration (angles and sides) of both triangles is obviously identical. Let t_1, t_2, t_3 be the position vectors of P_1, P_2, P_3 ; all are, of course, unit vectors. Similarly t'_1, t'_2, t'_3 are defined (Fig.36.2). Then

$$t'_1 = R_1 t_1, \quad t'_2 = R_1 t_2, \quad t'_3 = R_1 t_3. \quad (36-17)$$

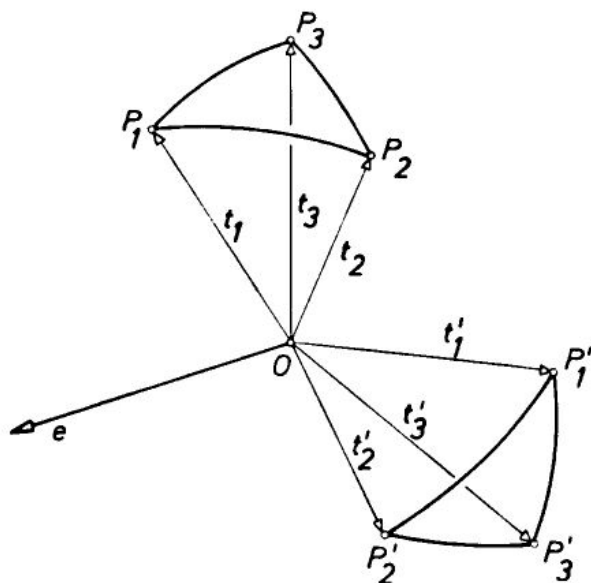


FIGURE 36.2. Rotation of a configuration.

Each of the vectors t_1, t_2, t_3 ; t'_1, t'_2, t'_3 can be obtained by rotating a fixed unit vector e (for which we may take, for instance, the unit vector of the X-axis) by a certain matrix $R(\omega_1), R(\omega_2), \dots, R(\omega'_3)$; we write $R(\omega_i)$ instead of R_{ω_i} to avoid two-level subscripts. Then, for $i = 1, 2, 3$,

$$t_i = R(\omega_i)e, \quad t'_i = R(\omega'_i)e. \quad (36-18)$$

Combining (36-17) and (36-18) we have

$$t'_1 = R(\omega'_1)e = R_1 R(\omega_1)e$$

or

$$R(\omega'_1) = R_1 R(\omega_1) . \quad (36-19)$$

Here $i = 1, 2, 3$, but clearly the configuration rotated by R_1 can have any number of points.

Thus, multiplying, *from the left*, a set of rotation matrices $R(\omega_i)$, $i = 1, 2, 3, \dots$, by a fixed matrix R_1 preserves the configuration. The geometrical configuration is invariant with respect to left multiplication. Similarly we may show the invariance of geometry with respect to right multiplication.

Homogeneity and isotropy imply that the essential properties depend only on the geometrical configuration. Therefore, also the probability measure must be invariant with respect to right and left multiplication. It can be proved that for a compact group such as the rotation group, there is essentially (apart from a constant factor) only one group measure that is both right and left invariant (Smirnow, 1964a, § 89). Such an invariant volume element in rotation group space may be shown to be

$$dV = \sin\theta d\theta d\Lambda dA \quad (36-20)$$

(Moritz, 1978c, sec.6). The total volume of group space is, by (36-3),

$$V = \int_{\Lambda=0}^{2\pi} \int_{\theta=0}^{\pi} \int_{A=0}^{2\pi} \sin\theta d\theta d\Lambda dA = 8\pi^2 ,$$

so that

$$d\Omega = \frac{1}{8\pi^2} \sin\theta d\theta d\Lambda dA \quad (36-21)$$

is the desired element of probability measure.

Now we are ready to attack the computation of expectations and covariances. The expectation $E\{f(t, \omega)\}$ becomes

$$\begin{aligned}
E\{f(t, \omega)\} &= \iiint_{\Omega} f(R_{\omega} t) d\Omega \\
&= \iiint_{\Omega} f(R_{\omega} R_{-\tau} t) d\Omega \\
&= \iiint_{\Omega} f(R_{\omega} e_z) d\Omega ,
\end{aligned} \tag{36-22}$$

in view of right invariance and using (36-11). However, $R_{\omega} e_z$ transforms the unit vector e_z of the Z-axis into the unit vector of the z-axis, which has the spherical coordinates θ and Λ (Fig.36.1). Hence,

$$f(R_{\omega} e_z) = f(\theta, \Lambda) , \tag{36-23}$$

and (36-22) becomes

$$E\{f(t, \omega)\} = \frac{1}{8\pi^2} \int_{\Lambda=0}^{2\pi} \int_{\theta=0}^{\pi} \int_{\Lambda=0}^{2\pi} f(\theta, \Lambda) \sin\theta d\theta d\Lambda d\Lambda .$$

We integrate over Λ and replace θ, Λ by θ, λ , respectively; obviously, the symbols for the integration variables are irrelevant. The result is

$$E\{f(t, \omega)\} = \frac{1}{4\pi} \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} f(\theta, \lambda) \sin\theta d\theta d\lambda , \tag{36-24}$$

which is simply the average over $f(t)$ over the unit sphere. It is zero if $f(\theta, \lambda)$ contains no zero-degree spherical harmonic, which corresponds to our usual assumption that the stochastic process under consideration is centered.

Then, by (34-16), the covariance function becomes

$$C(t, u) = E\{f(t)f(u)\} \tag{36-25}$$

with

$$t = [\theta, \lambda] , \quad u = [\theta', \lambda'] . \tag{36-26}$$

More explicitly this is written

$$C(t, u) = \iiint_{\Omega} f(R_{\omega} t) f(R_{\omega} u) d\Omega , \tag{36-27}$$

which, because of right invariance, is equal to

$$c(t, u) = \int_{\Omega} f(R_{\omega} R_{-\tau} t) f(R_{\omega} R_{-\tau} u) d\omega. \quad (36-28)$$

Now, by (36-11) and (36-23),

$$f(R_{\omega} R_{-\tau} t) = f(\theta, \lambda) \quad (36-29)$$

gives the value of f at a point P with spherical coordinates (θ, λ) , whereas

$$f(R_{\omega} R_{-\tau} u) = f(\theta', \lambda') \quad (36-30)$$

denotes the value of f at some point $Q = (\theta', \lambda')$ situated at the spherical distance ψ from P (Fig. 36.3). That the spherical distance ψ between P and Q is equal to the spherical distance between the points t and u as given by (34-23) follows from the invariance of the configuration with respect to the rotation $R_{\omega} R_{-\tau}$. It is also easily seen that if α is chosen as the azimuth from t to u , then A in (36-2) will be the azimuth from P to Q .

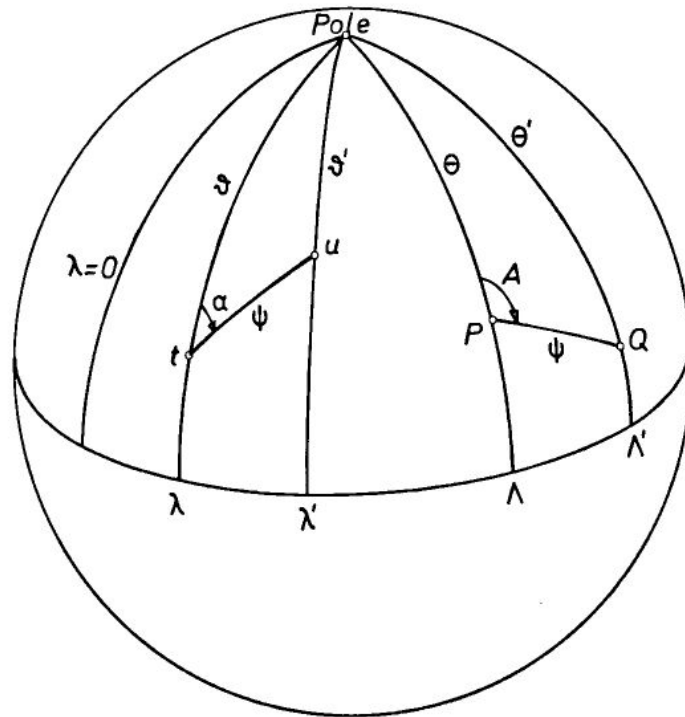


FIGURE 36.3. Invariance of the spherical distance ψ .

Thus (36-28) becomes

$$C(t,u) = \frac{1}{8\pi^2} \int_{\Lambda=0}^{2\pi} \int_{\Theta=0}^{\pi} \int_{\Lambda=0}^{2\pi} f(\Theta, \Lambda) f(\Theta', \Lambda') \sin \Theta d\Theta d\Lambda dA . \quad (36-31)$$

On replacing Θ, Λ, A by θ, λ, α we get

$$C(t,u) = \frac{1}{8\pi^2} \int_{\lambda=0}^{2\pi} \int_{\theta=0}^{\pi} \int_{\alpha=0}^{2\pi} f(\theta, \lambda) f(\theta', \lambda') \sin \theta d\theta d\lambda d\alpha . \quad (36-32)$$

Formally, this is only a change in the symbols for the integration variables; but geometrically the meaning is now altered profoundly: a comparison with (34-25) shows that this is simply the average M , so that by (34-26),

$$C(t,u) = E\{f(T, \omega) f(u, \omega)\} = M\{f(t) f(u)\} = r(\psi) ; \quad (36-33)$$

the true and the empirical covariance functions are identical.

Exactly as in the two-dimensional case, the reason for this identity is simply that probability space is made to coincide with rotation group space so that the expectation E and the rotation group average M are identical; our present model is *trivially ergodic*.

Commutativity of the operators L_i and M . In sec. 11 we have used the commutativity of the averaging operator M and the linear functionals L_i :

$$L_i M = M L_i . \quad (36-34)$$

This commutativity is not at all obvious if the definition (10-2) is used; take, as an example, a horizontal derivative

$$L_i = \frac{\partial}{\partial \theta} . \quad (36-35)$$

It is, however, clear for (36-25):

$$L_i^t C(t,u) = L_i^t E\{f(t) f(u)\} = E\{(L_i^t f(t)) f(u)\} . \quad (36-36)$$

The symbol L_i^t indicates that the functional L_i acts on the point t ; the notation is fully analogous to (11-12). For example, if L_i is a

differential operator, then (36-36) means simply the permissible differentiation of a definite integral with respect to a parameter under the integral sign, t being a parameter in the integral (36-27). Since M equals E by (36-33), also the commutativity (36-34) holds.

The total average \bar{E} . In sec. 14 we have defined the total average \bar{E} by (14-12) as

$$\bar{E} = EM,$$

(36-37)

that is, as an average both over the probabilistic distribution of the noise n (expectation E) and over the rotation group space (mean M). This can now be explained as follows.

We interpret the rotation group space as a formal probability space for the signal, that is, for the random functions $f(t, \omega)$. Then the probability space for the whole system, consisting of signal and noise, may be taken as the product space (cartesian product) of the rotation group space and the probability space of the noise. This implies the definition (36-37), the statistical independence of signal and noise, and hence relations (14-11) and (14-20).

37. STATISTICAL DISTRIBUTIONS IN ROTATION GROUP SPACE

In the last section we have studied rotation group space as a formal probability space primarily with respect to covariances. The covariance theory of stochastic process is what is needed for linear least squares prediction and estimation problems; it can be treated without explicit reference to the underlying statistical distributions, of which only the moments of first order (mean values), and of second order (variances and covariances) are needed.

Even in least-squares prediction and collocation, however, the distributions of relevant quantities are required if we wish to perform *statistical tests*. Already to answer very elementary but meaningful questions we need distributions.

Such a question is, for instance: What is the average global relative frequency of a $1^\circ \times 1^\circ$ mean gravity anomaly situated between -28 and -36 mgal? This question is answered by the histogram of Fig. 37.1; the relative frequency is the number of $1^\circ \times 1^\circ$ anomalies having magnitude within a specified interval, divided by the total number of $1^\circ \times 1^\circ$ anomalies.

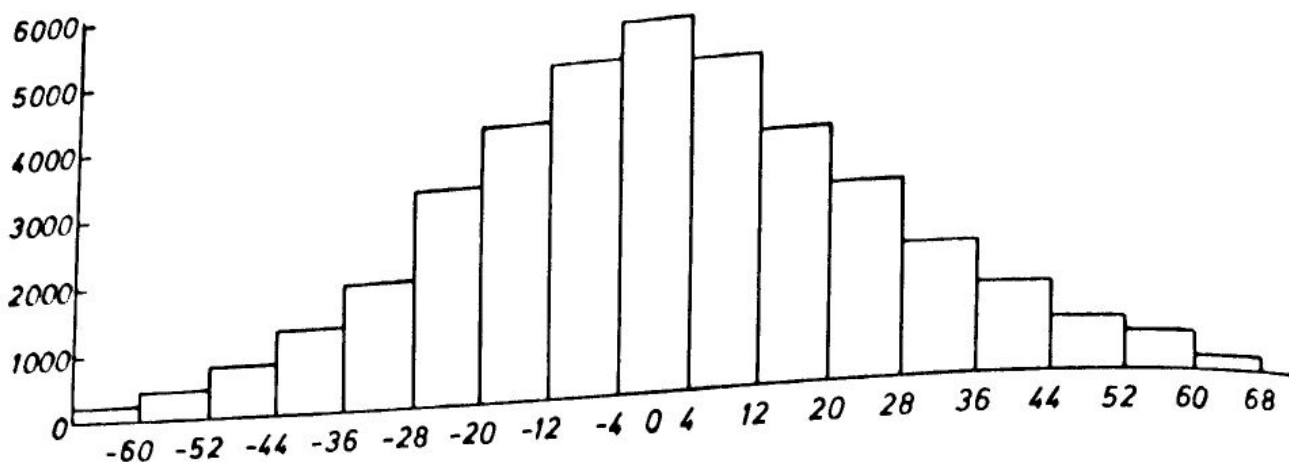


FIGURE 37.1. Number of $1^\circ \times 1^\circ$ mean gravity anomalies having magnitude within a specified interval. After (Rapp, 1977, p.5).

A histogram looks very similar to a probability density curve. In fact, probability is conceptually closely related to relative frequency, and the relative frequency satisfies the axioms of probability theory; cf. (Gnedenko, 1967, sec.I.7).

In this sense, a relative frequency may be mathematically treated as a probability, even if the underlying phenomenon is not "stochastic" in some physical sense. Thus, we may consider the relative frequency, mentioned above, as a measure of the probability that the mean anomaly (supposed unknown) of a certain $1^\circ \times 1^\circ$ block lies between -28 and -36 mgal.

Thus, basically we have simply a problem of classification (of gravity anomalies according to size), but probability terminology is again convenient, as it was in the case of covariance functions.

Another relevant question would be, for instance: What is the probability (in the sense of relative frequency) that a $1^\circ \times 1^\circ$ mean Δg -value lies between -28 and -36 mgal and that the mean value of the geoidal height N for the same $1^\circ \times 1^\circ$ block lies between 25 and 30 meters?

To answer such and related questions, we must construct appropriate distribution functions for Δg , for Δg and N jointly, etc. The distribution density for Δg will be a continuous analogue of the histogram of Fig. 37.1. To find a suitable formal probability space, let us note that the number of relevant $1^\circ \times 1^\circ$ mean anomalies is counted regardless of the position, on the earth's surface, of the $1^\circ \times 1^\circ$ blocks under consideration. All positions on the sphere are treated equally: again we have

homogeneity and isotropy. Thus, rotation group space is seen to be the proper probability space also as a basis for the mathematical description of statistical distributions.

The nature of a statistical distribution is best illustrated by the case of a function of one variable. Therefore, we shall again start with the group of rotations of the circle (the two-dimensional rotation group), which is parametrized by one variable ω , $0 \leq \omega < 2\pi$. Probability space is, therefore, the unit circle, and the element of probability measure is $\frac{1}{2\pi} d\omega$, which is clearly left and right invariant. Let the random function under consideration be denoted by $f(\omega)$.

We plot ω along the horizontal axis of a graph; then $f(\omega)$ is defined for $0 \leq \omega < 2\pi$ (it could, of course, be continued periodically for other abscissas).

Then the distribution function $\Phi(x)$ is defined by

$$\Phi(x) = \text{Prob}\{f(\omega) < x\} \quad (37-1)$$

as the probability that $f(\omega)$ takes a value smaller than x . It is the measure of all values of ω for which $f(\omega) < x$; this measure is obviously a function of x . In the situation shown in Fig. 37.2, $f(\omega) < x$ if ω is contained in the interval AB or in the interval CD ; thus

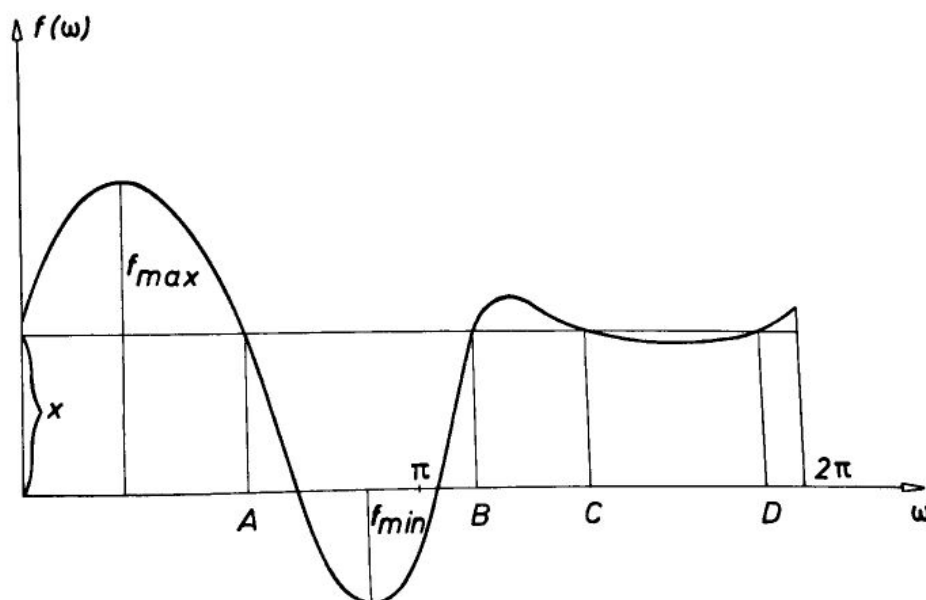


FIGURE 37.2. Distribution of a random function $f(\omega)$.

$$\text{Prob}\{f(\omega) < x\} = \frac{1}{2\pi}(\overline{AB} + \overline{CD}) ,$$

\overline{AB} denoting the length of AB and the factor $1/2\pi$ serving to make the measure of the total interval from 0 to 2π equal to unity.

Generally we may write

$$\Phi(x) = \frac{1}{2\pi} \int_{f(\omega) < x} d\omega , \quad (37-2)$$

the integral being extended over those points ω for which $f(\omega) < x$.

The derivative of (37-1) with respect to x gives the probability density

$$\phi(x) = \Phi'(x) = \frac{d\Phi}{dx} , \quad (37-3)$$

which has an even more intuitive geometrical meaning. For a differential dx (operations with differentials are justified in the usual way) we have

$$\phi(x)dx = d\Phi(x) = \text{Prob}\{x < f(\omega) < x + dx\} . \quad (37-4)$$

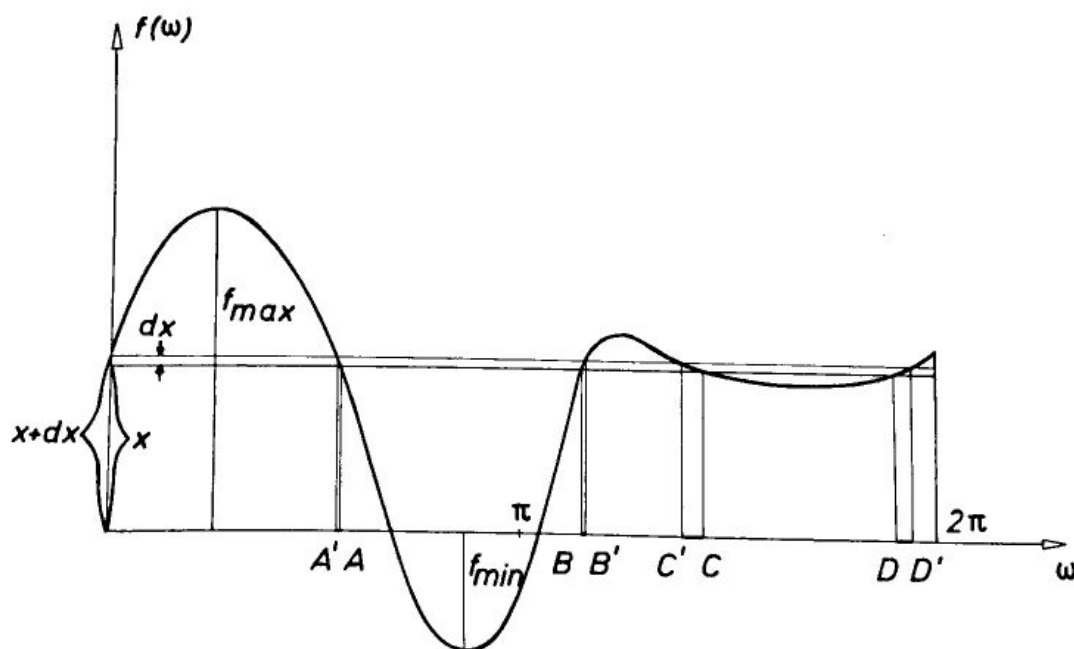


FIGURE 37.3. Definition of the distribution density.

According to Fig. 37.3, which represents the same function, $f(\omega)$ assumes a value between x and $x+dx$ if x lies in one of the small intervals $A'A$, BB' , $C'C$, or DD' , so that

$$d\phi(x) = \frac{1}{2\pi} (A'A + BB' + C'C + DD') , \quad (37-5)$$

or

$$\phi(x) = \frac{1}{2\pi dx} (A'A + BB' + C'C + DD') , \quad (37-6)$$

which gives an intuitive geometrical interpretation of the distribution density $\phi(x)$.

The distribution function $\phi(x)$ is a monotone nonnegative function of x , $-\infty < x < \infty$, as shown in Fig. 37.4. It is identically zero for $x < f_{\min}$ (there is no ω for which $f(\omega) < f_{\min}$) and identically one for $x > f_{\max}$ (for all ω there is $f(\omega) < x$ if $x > f_{\max}$).

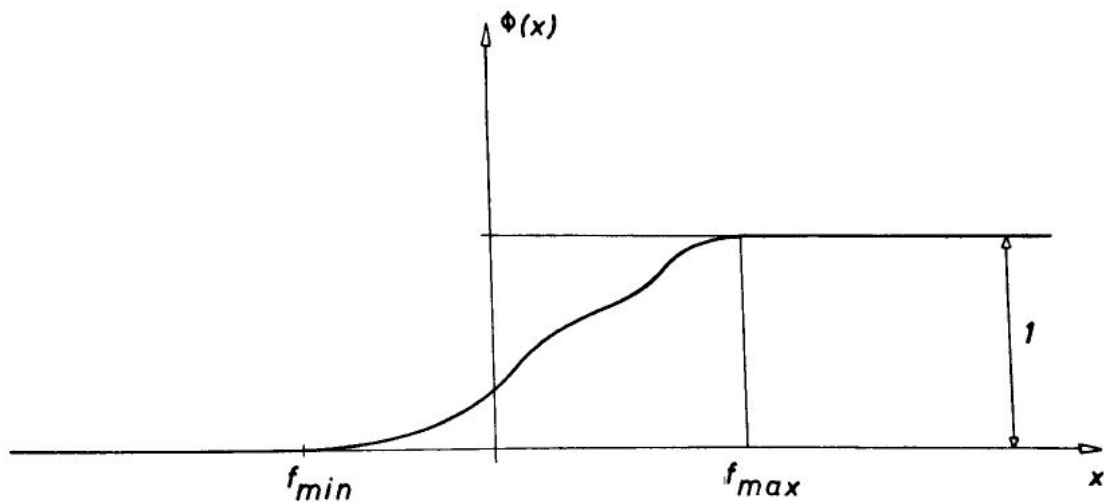


FIGURE 37.4. A distribution function.

The statistical expectation E of the random variable f can now be computed in two ways: by means of the probability measure $d\omega/2\pi$:

$$E\{f\} = \frac{1}{2\pi} \int_0^{2\pi} f(\omega) d\omega , \quad (37-7)$$

or by means of the distribution function

$$E(f) = \int_{-\infty}^{\infty} x d\Phi(x) . \quad (37-8)$$

Geometrically, both expressions give the area under the curve $f(\omega)$ in Fig. 37.2; (37-7) corresponds to the Riemann partitioning and (37-8) corresponds to the Lebesgue partitioning of the same definite integral; therefore (37-7) and (37-8) are identical; cf. (Feller, 1966, p.115-116) and (Kolmogorov and Fomin, 1970, p.293).

Let us, finally, consider the stochastic process (33-12),

$$f(t, \omega) = f(t + \omega) . \quad (37-9)$$

In view of the rotational invariance, the distribution function of $f(t + \omega)$, for fixed t , is the same for all t , hence, is the same as for $t = 0$:

$$\text{Prob}\{f(t + \omega) < x\} = \text{Prob}\{f(\omega) < x\} . \quad (37-10)$$

This is immediately seen from Fig. 37.2: replacing t by $t + \omega$ means only a translation of the figure as a whole to the right or left, the length of the intervals AB and CD remaining unchanged.

What is more, we may also write

$$\Phi(x) = \text{Meas}\{f(t) < x\} , \quad (37-11)$$

using only the sample function $f(t)$ defined on the "space circle" $0 \leq t < 2\pi$ with measure "Meas" defined by its element $dt/2\pi$, without any probabilistic interpretation; this follows immediately by replacing ω in (37-1) by t . This is formally very simple, but conceptually of fundamental importance: it shows that we may consistently work with one basic sample function $f(t)$ only and still avail ourselves of the formal advantages of probability theory.

Thus (37-4) is now written as

$$d\Phi(x) = \Phi'(x)dx = \text{Meas}\{x < f(t) < x + dx\} ,$$

which clearly shows that our formal "probability" is really nothing else than a relative frequency.

Distributions in three-dimensional rotation group space. After these preliminaries we come to the geodetically relevant case of three-dimensional rotations. The basic ideas remain the same, though the notation is more cumbersome.

Let $f(\omega)$ be a real-valued random function; the argument is defined by (36-2). Then the distribution function $\phi(x)$ of f is

$$\phi(x) = \text{Prob}\{f(\omega) < x\} . \quad (37-12)$$

It should be noted that x is a one-dimensional real variable, $-\infty < x < \infty$, though ω denotes a point in three-dimensional rotation group space Ω . Consider now the random function (36-13),

$$f(t, \omega) = f(R_\omega t) \quad \text{with} \quad t = [\theta, \lambda] . \quad (37-13)$$

In view of the rotational invariance, the distribution function of

$$\phi(x) = \text{Prob}\{f(R_\omega t) < x\} \quad (37-14)$$

does not depend on t . Following the reasoning that leads from (36-22) to (36-24) we find that

$$\phi(x) = \text{Meas}\{f(\theta, \lambda) < x\} . \quad (37-15)$$

The measure "Meas" is surface measure on the unit sphere, normalized by the factor $1/4\pi$; its element is, as usual,

$$\frac{1}{4\pi} d\sigma = \frac{1}{4\pi} \sin\theta d\theta d\lambda . \quad (37-16)$$

Just as in (36-24), there is no longer an explicit dependence on the azimuth variable λ .

For the distribution density

$$\phi(x) = \phi'(x) \quad (37-17)$$

we have again a geometrical interpretation (Fig.37.5). Draw the neighboring contour lines

$$f(\theta, \lambda) = x = \text{const.},$$

$$f(\theta, \lambda) = x + dx = \text{const.}$$

on the sphere; they will, in general, consist of several unconnected closed curves. Let the areas between these neighboring closed lines be denoted by

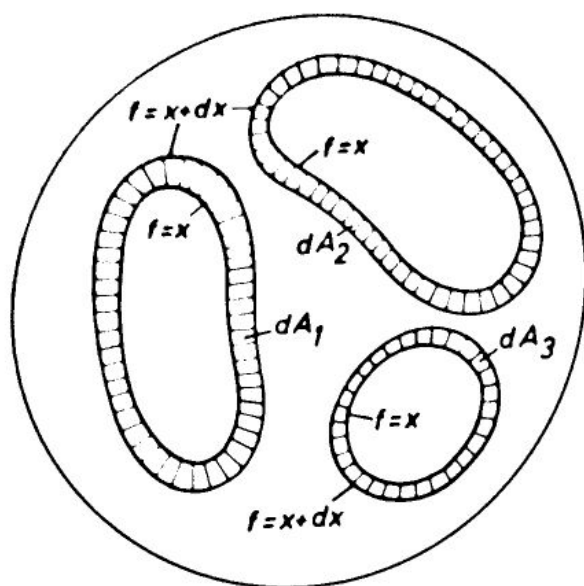


FIGURE 37.5. Geometrical interpretation of the distribution density.

dA_1, dA_2, dA_3, \dots (hatched in Fig. 37.5). Then

$$\phi(x)dx = \frac{1}{4\pi}(dA_1 + dA_2 + dA_3 + \dots) . \quad (37-18)$$

The distribution function $\phi(x)$ itself can be expressed in a form analogous to (37-2):

$$\phi(x) = \frac{1}{4\pi} \iint_{f(\theta, \lambda) < x} \sin \theta d\theta d\lambda . \quad (37-19)$$

Another basic problem is the determination of the joint distribution of two functions f and g on the sphere, say, of

$$\begin{aligned} f(\theta, \lambda) &= T(\theta, \lambda) , \\ g(\theta, \lambda) &= \Delta g(\theta, \lambda) , \end{aligned} \quad (37-20)$$

T and Δg denoting the disturbing potential and the gravity anomaly, respectively. The joint distribution function is

$$\phi(x, y) = \text{Meas}\{f(\theta, \lambda) < x, g(\theta, \lambda) < y\} . \quad (37-21)$$

The corresponding density is

$$\phi(x,y) = \frac{\partial^2 \phi(x,y)}{\partial x \partial y} ; \quad (37-22)$$

it may be geometrically illustrated as follows. Draw the contour lines

$$\begin{aligned} f(\theta, \lambda) &= x , \\ f(\theta, \lambda) &= x + dx , \end{aligned} \quad (37-23)$$

as well as the contour lines

$$\begin{aligned} g(\theta, \lambda) &= y , \\ g(\theta, \lambda) &= y + dy \end{aligned} \quad (37-24)$$

(Fig.37.6). The ribbons formed in this way intersect in areas dA_1, dA_2, dA_3, \dots (hatched in Fig.37.6), and

$$\phi(x,y)dx dy = \frac{1}{4\pi}(dA_1 + dA_2 + dA_3 + \dots) . \quad (37-25)$$

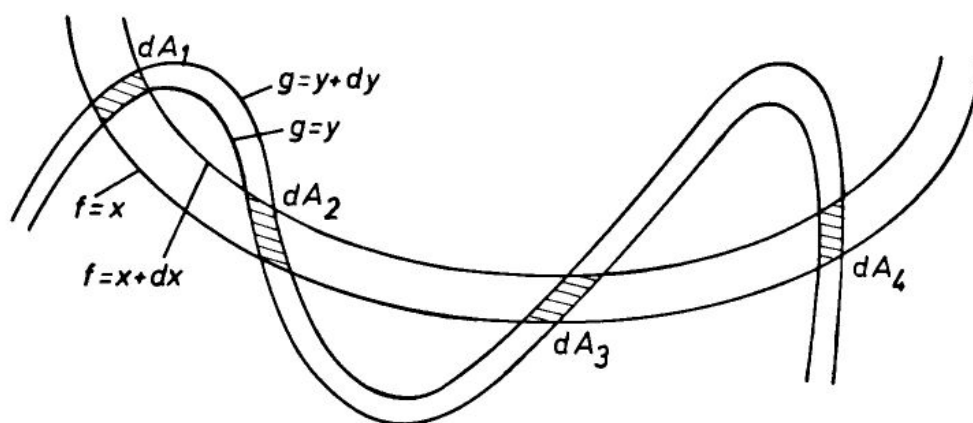


FIGURE 37.6. Joint distributions.

A final example will indicate how an azimuth-dependent situation can be handled. Consider the problem of the joint distribution of gravity anomalies at two points that are at a spherical distance ψ apart:

$$\Phi(x,y) = \text{Prob}\{f(t,\omega) < x, f(u,\omega) < y\} \quad (37-26)$$

where

$$\psi = \text{angle}(t,u) = \text{const.} \quad (37-27)$$

We have

$$t = [\theta, \lambda] , \quad (37-28)$$

$$u = [\theta', \lambda'] , \quad (37-29)$$

where the condition (37-27) can be written in the form

$$\cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\lambda'-\lambda) = \cos\psi = \text{const.} \quad (37-30)$$

Then (37-26) can be expressed as the integral

$$\Phi(x,y) = \frac{1}{8\pi^2} \iiint_{B(x,y)} \sin\theta d\theta d\lambda d\alpha , \quad (37-31)$$

where the integration is extended over the region $B(x,y)$ defined by the inequalities

$$\begin{aligned} f(\theta, \lambda) &< x , \\ g(\theta', \lambda') &< y ; \end{aligned} \quad (37-32)$$

θ', λ' are expressed as functions of $(\theta, \lambda, \alpha)$ by the trigonometric relations

$$\begin{aligned} \cos\theta' &= \cos\theta\cos\psi + \sin\theta\sin\psi\cos\alpha , \\ \sin(\lambda'-\lambda) &= \sin\psi\sin\alpha/\sin\theta' , \end{aligned} \quad (37-33)$$

which follow from the spherical triangle of Fig. 37.7. The integral (37-31) is analogous to (37-2) and (37-19); the probability measure $\sin\theta d\theta d\lambda d\alpha$ has been replaced by "spatial" (surface plus azimuth) measure $\sin\theta d\theta d\lambda d\alpha$ in the same way as (36-31) has been replaced by (36-32). Again we see that formal "probabilities" are really relative frequencies.

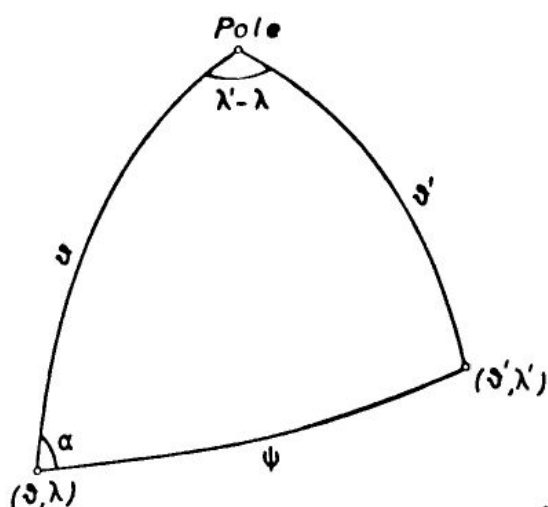


FIGURE 37.7. The basic spherical triangle.

These three examples illustrate the basic principles of the determination of single and joint distributions. Other cases can be handled similarly. In any case, we can operate with "spatial" functions $f(\theta, \lambda)$, $g(\theta, \lambda)$, ... only.

In practice, the functions $f(\theta, \lambda)$ are usually represented by discrete mean values (say, $5' \times 5'$ or $1^\circ \times 1^\circ$ block averages), and the integrations are to be replaced by sums.

38. THE MEANING OF STATISTICS IN COLLOCATION

Are gravity anomalies a stochastic phenomenon? There are different answers to this question.

To get one answer, consider gravity at one observation station and observe it repeatedly. Assume that the measuring errors are negligibly small, and remove known geophysical effects, especially tidal ones. Then the results of measurements at different times will be practically constant. We conclude: gravity is a deterministic, not a stochastic phenomenon.

This way of looking at the problem, repeating the same experiment under identical conditions, is the way we look at random measuring errors and at many other "stochastic" physical phenomena, the classical case being the repeated throw of a die. If the experimental results vary randomly, then we have a genuinely stochastic phenomenon. Under the assumption that the outcomes of the repeated experiment are independent of one another, we have the scheme of repeated trials, fundamental in probability theory.

There is, however, also another way of looking at the question of stochasticity of gravity anomalies. They are caused by mass anomalies, visible and invisible ones. These mass anomalies show some regular features, for instance, mountain chains extending in a regular fashion from north to south. From a regional or global point of view, however, the pattern is still quite irregular.

In fact, the mass distribution in the crust results from a superposition of many causes, such as tectonic and igneous activity and the action of water, ice, and wind. The resulting effect of these causes is irregular indeed and may well be considered random.

This is particularly true if we remove known features or trends, for instance, by a topographic-isostatic reduction to be mentioned below.

In this sense, the gravity anomalies, which are caused by such mass anomalies, may also be said to be random. The randomness exists here not with respect to time, as it was in the first case (measurement at the same point but at different times) but with respect to space (measurement at the same time but at different points). The random behavior is more or less independent of position on the sphere and of direction; it is homogeneous and isotropic.

Stochastic process interpretation. What do we mean by considering the anomalous gravitational potential T as a stochastic process? In conformity with sec. 34 we put

$$T = f(\theta, \lambda; \omega) . \quad (38-1)$$

The potential T at sea level is thus considered as a function of the spherical coordinates θ and λ (we employ the spherical approximation as usual), as well as of a phase variable ω which is a point in a probability space Ω . To each choice of ω there corresponds a function of θ and λ . The actual terrestrial potential T would then correspond to one particular choice, say ω_1 .

This means that the mathematical model comprises not only one earth, our Earth, but a number of other fictitious "sample earths", each of which corresponds to a different point $\in \Omega$.

Such a model is not a priori impossible (similar models are employed, e.g., in statistical thermodynamics); it may very well be used if there are good reasons: if it is mathematically convenient and provides practically meaningful results.

In the present geodetic case the use of such a model could perhaps be justified on mathematical grounds if it were Gaussian and ergodic; however, this is impossible in view of Lauritzen's theorem (sec.35). Whether non-

Gaussian models such as the one given as the First Example in sec. 35 (p.286) are mathematically attractive enough to serve as a stochastic process model for the earth's gravitational field, remains a matter of taste. The present author prefers to look into another direction, which is indicated by the Second Example in sec. 35 (p.288).

Statistics without stochastics. Our principal reason for introducing stochastic processes has been to avail ourselves of the corresponding mathematical apparatus and the statistical terminology, such as the concept of covariance functions; this is very convenient, useful, and of considerable heuristic value. If it is possible to retain this apparatus while at the same time working with one Earth only, then there is little reason why we should not take "Occam's razor" and cut away all the other fictitious "sample earths". (Occam: "*Entia non sunt multiplicanda praeter necessitatem*".)

The situation may well be compared to the global statistics of the human population at a given time. There are random variations from one human individual to the other. We have regular trends, such as the racial and cultural background, but there are genuinely irregular features left, distributed over the human population and thus over the earth's surface. This is not completely unlike the surface distribution of gravity anomalies (although the analogy, if pushed too hard, quickly becomes nonsense).

Is it permitted to study the global population statistics at a given time, to calculate various statistical distributions? Everyone will answer this question in the affirmative, although there is only *one* global population. All statistical distributions are simply calculated on the basis of this population.

This is the subject of *descriptive statistics*, which computes relative frequencies, histograms, distribution functions, mean values, variances, and covariances only on the basis of the available population (cf. Kreyszig, 1970, Part I; Mises, 1957, pp.166-167). This has the character of a classification of the data and does not necessarily presuppose a "stochastic" behaviour.

It appears that the statistics of the gravitational field should be handled similarly. We simply must take seriously the fact that there is only one gravitational field, and compute the whole statistics from this one field only.

The appropriate mathematical apparatus for studying the "second-order statistics" (variances and covariances) of the gravitational field is thus Norbert Wiener's (1930) "covariance analysis of individual functions" (Doob, 1949, sec.1). This model is implicit in almost all geodetic work in this field (Kaula, 1959, 1967; Heiskanen and Moritz, 1967, chapter 7); explicitly it was formulated in (Moritz, 1973a, sec.8). It essentially uses

the idea of homogeneity and isotropy. For the sphere, homogeneity and isotropy really form a single compound notion, namely, invariance under rotations (in contrast to the plane, where homogeneity, invariance under translation, and isotropy, invariance under rotation, are separate notions!); this motivates the introduction of the three-dimensional rotation group.

Throughout the present book we have used such a covariance analysis of individual functions applied to the anomalous potential T , beginning with sec. 10. In sec. 37 we have extended the idea of a statistical analysis of the individual function T to the computation of statistical distributions of T and related quantities.

It is now of basic importance that, *formally*, this theory can also be interpreted within the framework of stochastic processes, as our ergodic Second Model (p.288). This is, of course, independent of the question whether the anomalous gravitational field is "really" a stochastic process in some physical sense. Probability theory then simply serves to provide a convenient mathematical formalism.

The deeper reason why this formalism can be applied is that, mathematically, both probabilities and relative frequencies satisfy the axioms of *measure theory*. Therefore, if we wish to avoid the term "probability", we might simply speak of "measure". However, this concept is rather abstract, whereas probabilistic terminology possesses an attractive intuitive flavor and is thus frequently preferred.

The problem whether and in which respect the anomalous gravitational field is a "genuinely stochastic phenomenon" will be answered differently by different people, depending on their scientific outlooks. Even with respect to the philosophical meaning of "probability" and "stochastic phenomenon" there are many different, even quite opposite, opinions, as is seen by comparing books such as (Gnedenko, 1967), (Finetti, 1974), and (Mises, 1957). Still, the mathematical formalism is the same.

Similarly, the mathematical formalism, proposed here as a statistical background of collocation, is independent of how seriously we take the stochastic character of the gravitational field. Even if we rigorously deny this stochasticity, we can still accept the formal statistical analysis presented here: we then have "statistics without stochastics"; cf. (Sansô, 1978c).

Randomness again. At any rate, the standard errors computed as rotation group means are meaningful as global averages. Consider, for instance, least-squares gravity interpolation or prediction, as mentioned on pp.80-81. If we use a global covariance function, then the standard error of prediction computed by a formula of type (9-29) will characterize the average interpolation accuracy on a global scale. This does not necessarily mean

that it will also give the average accuracy in a certain region. The latter accuracy would be better obtained by using a regional covariance function, which characterizes the statistical behavior of the gravity anomalies in that particular region.

For global data combination problems it is usually the global average accuracy which we need; the present book concerns itself primarily with those problems. Still, regional averages are also of interest, and it would be desirable if the statistical characteristics in different regions were similar to each other and to the corresponding global characteristics.

This would be the case if averages over a certain region (including an averaging over different azimuths) were approximately equal to the global rotation group average. A comparable situation would be throwing a die a large number of times. If the average frequency of throwing a "3" in any sample, say, of a hundred throws (e.g., the average of throws No. 1 to No. 100, the average of throws No. 101 to 200, of No. 201 to 300, and so on) is approximately equal, about $1/6$, then we say that the die behaves in a properly random fashion. This is the much quoted stability of relative frequency, which is considered characteristic for stochastic phenomena; cf. (Gnedenko, 1967, sec.I.7).

Thus we might similarly say that the gravity anomalies behave randomly if regional averages (of not too small a size) are approximately equal to each other.

Obviously this is not in general the case. Statistical characteristics will be different in a mountainous region, in a plane region, and in an oceanic region, even if they have the same size. These characteristics may also depend on the azimuth--this is called anisotropy--, for instance if we have a large mountain chain extending essentially in one direction such as the Rocky Mountains. We then speak of regional trends. We may try to remove these trends, for instance by gravity reduction to be discussed below. In this way we may hope to get a more "random" behavior of the residual gravity anomalies, more independent of position and azimuth, rendering various regional averages similar to each other and to the global average.

Another possibility which might be used for local applications is to consider local or, exceptionally, anisotropic covariance functions; see, for instance, (Groten et al., 1979), (Kearsley, 1977), and (Morrison, 1977).

For global applications, however, the rotation group average, which is homogeneous and isotropic by definition, seems to be the appropriate concept.

Gravity reduction as trend removal. From a statistical point of view it would be desirable to eliminate all known trends. For instance, an elimination of the correlation of the gravity anomalies with elevation is closely

related to the Bouguer reduction (cf. Heiskanen and Moritz, 1967, sec.7-10). In this way, the very *local* topographic effects are removed and the gravity anomalies become much smoother and easier to interpolate. On the other hand, the Bouguer anomalies are approximately proportional to the average height of an area and thus have the character of a *regional* trend. The removal of this latter trend, in addition to the local one, is effected by an isostatic reduction.

If Δg is the given free-air gravity anomaly, then

$$\Delta g_B = \Delta g - 2\pi G \rho h \quad (38-2)$$

(G ... gravitational constant, ρ ... density, h ... topographic elevation) is an approximate expression for the Bouguer anomaly, which is roughly equal to

$$- 2\pi G \rho h_m$$

(h_m ... mean elevation of an area). Therefore, the isostatic anomaly, which is approximately given by

$$\Delta g_I = \Delta g - 2\pi G \rho (h - h_m) \quad (38-3)$$

(Moritz, 1968b, sec.6), can be said to be free of the principal trends (local elevation h and mean elevation h_m); it can be expected to be smaller, smoother and more "random" than the free-air anomaly. In fact, if isostatic compensation were perfect, then Δg_I would be zero.

The great practical importance of the Bouguer reduction for gravity interpolation and of topographic-isostatic reduction for interpolation of deflections of the vertical is well known. Thus, for local or regional purposes, such reductions are of great value.

In global data combinations, a consistent isostatic reduction of all data under consideration, and the subsequent "antireduction" of the results obtained, would certainly improve the results of least-squares collocation. Unfortunately, this improvement has to be paid for by an enormous computational effort, which may be justified in some cases but probably not in general.

It should, however, be mentioned that the gross features of an isostatic reduction can be calculated relatively easily from a spherical-harmonic expansion of topography; cf. (Lachapelle, 1976).

Removal of a lower-degree harmonic field. Another possibility of a trend removal is the subtraction of one of the available truncated spherical-

harmonic expansions such as found in (Lerch et al., 1977), (Papp, 1977) or (Gaposchkin, 1979). Such a field represents very well the regional features; it can be expected that the residual field after subtracting such a regional field will be smaller and more irregular than the original anomalous gravitational field.

The procedure is conceptually similar to the isostatic reduction:

1. "reduction" of the data by subtracting the effect of the regional field;
2. application of the collocation formulas to the residuals;
3. "antireduction" of the computed quantities by adding the corresponding effect of the regional field.

The computational effort, however, is much less.

It is also possible to combine these two types of reduction; cf. (Lachapelle, 1975).

In principle one could also think of removing other trends, for instance, known density anomalies. This corresponds to the principle that all possible trends should be eliminated before applying statistics. As many other principles, this one is good as long as it is combined with common sense and not pushed to the extreme. The analytical side of collocation may help to find the proper measure: even if there is no trend removal whatsoever, collocation works as an analytical approximation method.

Group-invariant estimation. Our statistical interpretation makes essential use of the invariance with respect to the rotation group (sec.36). By a slight shift of perspective it is possible to stress this invariance, downplaying the statistical aspect. In this way we may interpret least-squares collocation as a rotation group invariant linear estimation (Sansó, 1978a).

Nonlinear estimation. The non-Gaussian character of the anomalous gravity field implies that the "best linear" estimate is not the absolutely best estimate: nonlinear estimation may still reduce the variance of the result. Therefore, nonlinear prediction has been suggested, among others, by Kaula (1966) and Grafarend (1972). In practice, linear methods are used almost exclusively.

The many facets of collocation. We have seen that least-squares collocation has its roots in many fields such as: (1) least-squares estimation, (2) stochastic processes, (3) approximation theory, (4) functional analysis, (5) potential theory, (6) inverse and improperly posed problems, and (7) group theory. Some of these relationships, especially the statistical and the analytical aspect, are interesting and significant enough to invite concentration on one or the other. However, a balanced and complete understanding of least-squares collocation involves a careful synthesis of all relevant aspects.

39. ELLIPSOIDAL CORRECTIONS

The quantities of the anomalous gravity field, such as gravity anomalies or geoidal heights, are relatively small. In formulas relating these quantities one usually neglects, therefore, the flattening of the reference ellipsoid, arriving at expressions which are formally spherical. This spherical approximation underlies almost all formulas in physical geodesy, such as Stokes' formula (sec.2) or Molodensky's series (sec.43); also the present treatment of least-squares collocation has been based on it.

This spherical approximation causes an error which is negligible in most practical applications. Lelgemann (1970, p.20) has shown that, in the case of Stokes' formula, this error is on a global average on the order of $\pm 0.2\text{m}$ in the geoidal height. This is one order of magnitude smaller than the accuracy implied by the present gravimetric and satellite data.

For higher accuracies, however, ellipsoidal corrections must be taken into account. This can be done in the following way, which is basically due to Sastrebin (1956) and has been used since by many authors.

Any quantity F of the anomalous gravity field (the anomalous potential, the geoidal height, the gravity anomaly, etc.) is expanded as a series with respect to a small parameter ϵ characterizing the deviation of the reference ellipsoid from a sphere:

$$F = F^{(0)} + \epsilon F^{(1)} + \epsilon^2 F^{(2)} + \epsilon^3 F^{(3)} + \dots \quad (39-1)$$

This parameter ϵ may be the flattening f or some similar ellipsoidal parameter; we shall take the square of the first excentricity:

$$\epsilon = e^2 = \frac{a^2 - b^2}{a^2} ; \quad (39-2)$$

for notations see sec. 2. In view of the smallness of the quantities under consideration, squares and higher powers of ϵ may be safely neglected, so that (39-1) reduces to

$$F = F^0 + e^2 F^1 ; \quad (39-3)$$

we have written $F^{(0)} = F^0$ and $F^{(1)} = F^1$. In this form we shall try to represent every quantity of the anomalous gravitational field.

Such a quantity also depends on the position as expressed, e.g., by the geodetic coordinates ϕ (latitude), λ (longitude), and h (height above the ellipsoid). If the quantity refers to the earth's surface, then h is a function of ϕ and λ , so that the quantity under consideration depends on ϕ and λ only. In this case, (39-3) may be written more explicitly:

$$F(\phi, \lambda) = F^0(\phi, \lambda) + e^2 F^1(\phi, \lambda) . \quad (39-4)$$

Since the function $F^0(\phi, \lambda)$ corresponds to $e = 0$, it may be considered as a function on a suitably defined sphere, e.g., a mean terrestrial sphere of radius

$$R = \sqrt[3]{a^2 b} \approx 6371 \text{ km} , \quad (39-5)$$

and ϕ, λ may be considered as spherical coordinates on this sphere.

In this way we have established a one-to-one mapping of the reference ellipsoid on a sphere of radius R by mapping a point of geodetic (geographical) coordinates (ϕ, λ) on the ellipsoid into a point of spherical coordinates (ϕ, λ) on the sphere; the values of ϕ and λ being numerically the same for both points.

After having defined, in this way, a mapping of points, we shall establish a mapping of functions by letting the function $F^0(\phi, \lambda)$ on the sphere correspond to the function $F(\phi, \lambda)$ on the ellipsoid, the two functions being related by (39-4).

The functions $F^1(\phi, \lambda)$ are determined as follows. For the anomalous potential T , we define

$$T^1(\phi, \lambda) = 0 , \quad (39-6)$$

so that we have the fundamental mapping equation:

$$T^0(\phi, \lambda) = T(\phi, \lambda) . \quad (39-7)$$

Then geoidal height, deflection of the vertical, gravity anomalies and similar quantities $F^0(\phi, \lambda)$ on the sphere are uniquely defined in terms of the basic function $T^0(\phi, \lambda)$ by spherical relations. On the other hand, the corresponding functions $F(\phi, \lambda)$ represent the actual values of these quantities on the ellipsoid and are likewise uniquely related to $T(\phi, \lambda) = T^0(\phi, \lambda)$ by appropriate ellipsoidal formulas. Thus the functions $F^1(\phi, \lambda)$

are likewise well defined; their determination is the purpose of the present section.

Now we are also in a position to better understand the meaning of the spherical approximation. It is the mapping of an ellipsoidal point with geographical (geodetic) coordinates (ϕ, λ) into a point of the sphere whose spherical coordinates are numerically equal to the coordinates (ϕ, λ) of the ellipsoidal point; furthermore all first-order quantities $e^2 F^1$ and terms of higher order are neglected.

In the present situation we have the same mapping by numerically identifying the geodetic coordinates on ellipsoid and sphere; however, now the first-order quantities are retained.

It should be mentioned that the mapping of geodetic coordinates into spherical coordinates is by no means the only way of establishing a correspondence between the reference ellipsoid and an auxiliary sphere. One may also map the ellipsoidal coordinates consisting of reduced latitude β and longitude λ into spherical coordinates by identifying their respective numerical values. This has been done, e.g., by Sagrebin (1956) and Hotine (1969, pp.320-322). Employing β and λ leads to somewhat simpler formulas, but the use of ϕ and λ seems to be preferable from a practical point of view: it is laborious to transform geographical latitudes ϕ into reduced latitudes β for all points used, and the division of blocks for mean values of gravity, etc., corresponds to ϕ, λ and not to β, λ .

We also mention that Molodensky (Molodenskii et al., 1962) and Koch (1968), in their solutions of ellipsoidal boundary-value problems, map geocentric coordinates on the ellipsoid (geocentric latitude and longitude) into spherical coordinates.

Geographical coordinates have been used by Lelgemann (1970, 1973) and Moritz (1974).

Geoidal heights. The geoidal height N is related to the anomalous potential T by

$$N = \frac{T}{\gamma} \quad (39-8)$$

where γ is normal gravity at the ellipsoid.

By eq. (2-96) of (Heiskanen and Moritz, 1967) we have, approximately,

$$\gamma = \gamma_a (1 + f^* \sin^2 \phi) \quad (39-9)$$

where γ_a is normal gravity at the equator and

$$f^* = -f + \frac{5}{2}m \quad (39-10)$$

(*ibid.*, eq.(2-99)). To the same approximation we have

$$f = \frac{1}{2}e^2 ; \quad (39-11)$$

and an inspection of numerical values (*ibid.*, p.80) shows that, approximately,

$$m = \frac{1}{2}e^2 . \quad (39-12)$$

Thus, (39-9) becomes

$$\gamma = \gamma_a \left(1 + \frac{3}{4}e^2 \sin^2 \phi \right) . \quad (39-13)$$

Introducing the Legendre polynomial

$$P_2(\sin \phi) = \frac{3}{2} \sin^2 \phi - \frac{1}{2} \quad (39-14)$$

we have

$$\gamma = \gamma_a \left[1 + \frac{1}{4}e^2 + \frac{1}{2}e^2 P_2(\sin \phi) \right] . \quad (39-15)$$

Since the mean value (over the sphere) of $P_2(\sin \phi)$ is zero, the mean value of γ becomes

$$\gamma^0 = \gamma_a \left(1 + \frac{1}{4}e^2 \right) , \quad (39-16)$$

so that (39-15) may be written as

$$\gamma = \gamma^0 \left[1 + \frac{1}{2}e^2 P_2(\sin \phi) \right]$$

or, by (39-14),

$$\gamma = \gamma^0 \left(1 - \frac{1}{4}e^2 + \frac{3}{4}e^2 \sin^2 \phi \right) . \quad (39-17)$$

Hence, (39-8) becomes

$$N = \frac{T}{\gamma^0} \left(1 + \frac{1}{4} e^2 - \frac{3}{4} e^2 \sin^2 \phi \right) . \quad (39-18)$$

Taking γ^0 as the (constant) spherical value corresponding to the ellipsoidal value γ , we have

$$N^0 = \frac{T^0}{\gamma^0} = \frac{T}{\gamma^0} \quad (39-19)$$

as the value of the geoidal height in the spherical approximation, so that

$$N = N^0 + e^2 N^1 \quad (39-20)$$

with

$$N^1 = \left(\frac{1}{4} - \frac{3}{4} \sin^2 \phi \right) N^0 . \quad (39-21)$$

Deflections of the vertical. The components of the deflection of the vertical are given by

$$\begin{aligned} \xi &= - \frac{\partial N}{\partial s_\phi} , \\ \eta &= - \frac{\partial N}{\partial s_\lambda} , \end{aligned} \quad (39-22)$$

as the derivatives of the geoidal height N along a north-south direction (line element ds_ϕ) and along an east-west direction (line element ds_λ). Generally, the line element of the ellipsoid is expressed by

$$ds^2 = \mu^2 d\phi^2 + \nu^2 \cos^2 \phi d\lambda^2 , \quad (39-23)$$

where μ and ν are the principal radii of curvature of the ellipsoid given by

$$\mu = \frac{c}{V^3} , \quad (39-24)$$

$$\nu = \frac{c}{V} , \quad (39-25)$$

where we have put

$$v = (1 + e'^2 \cos^2 \phi)^{\frac{1}{2}} \quad (39-26)$$

and used eqs. (2-3) and (2-4); these relations are found in any text on geometrical geodesy.

Thus,

$$\begin{aligned} ds_{\phi} &= v d\phi, \\ ds_{\lambda} &= v \cos \phi d\lambda, \end{aligned} \quad (39-27)$$

so that (39-22) becomes

$$\begin{aligned} \xi &= -\frac{1}{v} \frac{\partial N}{\partial \phi}, \\ \eta &= -\frac{1}{v \cos \phi} \frac{\partial N}{\partial \lambda}. \end{aligned} \quad (39-28)$$

In the spherical approximation we have

$$\begin{aligned} \xi^0 &= -\frac{1}{R} \frac{\partial N^0}{\partial \phi}, \\ \eta^0 &= -\frac{1}{R \cos \phi} \frac{\partial N^0}{\partial \lambda}. \end{aligned} \quad (39-29)$$

On expanding we have to order e^2 :

$$R = \sqrt[3]{a^2 b} = a \left(1 - \frac{1}{6} e^2 \right), \quad (39-30)$$

$$c = \frac{a^2}{b} = a \left(1 + \frac{1}{2} e^2 \right) = R \left(1 + \frac{2}{3} e^2 \right), \quad (39-31)$$

$$v = 1 + \frac{1}{2} e^2 \cos^2 \phi = 1 + \frac{1}{2} e^2 - \frac{1}{2} e^2 \sin^2 \phi, \quad (39-32)$$

$$\frac{1}{v} = \frac{1}{R} \left(1 + \frac{5}{6} e^2 - \frac{3}{2} e^2 \sin^2 \phi \right), \quad (39-33)$$

$$\frac{1}{v} = \frac{1}{R} \left(1 - \frac{1}{6} e^2 - \frac{1}{2} e^2 \sin^2 \phi \right). \quad (39-34)$$

By (39-20) and (39-21) we get

$$\begin{aligned} \frac{\partial N}{\partial \phi} = \frac{\partial N^0}{\partial \phi} + e^2 \frac{\partial N^1}{\partial \phi} = \frac{\partial N^0}{\partial \phi} + e^2 \left(\frac{1}{4} - \frac{3}{4} \sin^2 \phi \right) \frac{\partial N^0}{\partial \phi} - \\ - e^2 \cdot \frac{3}{2} \sin \phi \cos \phi \cdot N^0, \end{aligned} \quad (39-35)$$

and similarly,

$$\frac{\partial N}{\partial \lambda} = \frac{\partial N^0}{\partial \lambda} + e^2 \left(\frac{1}{4} - \frac{3}{4} \sin^2 \phi \right) \frac{\partial N^0}{\partial \lambda}. \quad (39-36)$$

Substituting these expressions, together with (39-33) and (39-34), into (39-28) and taking (39-29) into account we finally obtain

$$\xi = \xi^0 + e^2 \xi^1, \quad \eta = \eta^0 + e^2 \eta^1 \quad (39-37)$$

with

$$\xi^1 = \left(\frac{13}{12} - \frac{9}{4} \sin^2 \phi \right) \xi^0 + \frac{3}{2} \sin \phi \cos \phi \frac{N^0}{R}, \quad (39-38)$$

$$\eta^1 = \left(\frac{1}{12} - \frac{5}{4} \sin^2 \phi \right) \eta^0. \quad (39-39)$$

In the expressions for ξ^1 and η^1 , we may use for ξ^0, η^0, N^0 any approximate values, for instance, such values as obtained from a truncated spherical-harmonic expression, the reason being that $e^2 \xi^1$ and $e^2 \eta^1$ are very small.

Gravity anomalies. The relation between the anomalous potential T and the gravity anomaly Δg is essentially more complicated. In fact, from Sagrebin (1956) to Lelgemann (1970), this relation has been found by solving a boundary-value problem which is an extension of Stokes' problem (cf. p.15) to the ellipsoid. We shall here follow a simpler approach, which has been outlined by Hotine (1969, pp.320-322), developing it for geographical coordinates ϕ and λ .

In these coordinates, the potential T at the ellipsoidal surface may be expressed by a spherical-harmonic expansion,

$$T(\phi, \lambda) = \sum_{n=2}^{\infty} \sum_{m=0}^n P_{nm}(\sin \phi) (A_{nm} \cos m \lambda + B_{nm} \sin m \lambda). \quad (39-40)$$

P_{nm} being the usual Legendre functions. From geometrical geodesy it is known that geographical latitude ϕ and reduced latitude β are related by

$$\tan \beta = \frac{b}{a} \tan \phi, \quad (39-41)$$

from which we obtain by a series expansion with respect to e^2 , restricted to linear terms,

$$\beta = \phi - \frac{1}{2} e^2 \sin \phi \cos \phi, \quad \phi = \beta + \frac{1}{2} e^2 \sin \beta \cos \beta. \quad (39-42)$$

Hence, to the same approximation,

$$P_{nm}(\sin \phi) = P_{nm}(\sin \beta) + \frac{1}{2} e^2 \frac{dP_{nm}}{d\beta} \sin \beta \cos \beta. \quad (39-43)$$

The following formulas are standard for spherical harmonics (cf. Gradshteyn and Ryzhik, 1965, p.1005):

$$\frac{dP_{nm}(\sin \beta)}{d\beta} \cos \beta = (n+1) \sin \beta P_{nm} - (n-m+1) P_{n+1,m}, \quad (39-44)$$

$$\sin \beta P_{nm}(\sin \beta) = \frac{n-m+1}{2n+1} P_{n+1,m} + \frac{n+m}{2n+1} P_{n-1,m}, \quad (39-45)$$

whence

$$\sin \beta \cos \beta \frac{dP_{nm}}{d\beta} = a_{nm} P_{n+2,m} + b_{nm} P_{nm} + c_{nm} P_{n-2,m} \quad (39-46)$$

with

$$\begin{aligned} a_{nm} &= - \frac{n(n-m+1)(n-m+2)}{(2n+1)(2n+3)}, \\ b_{nm} &= \frac{n^2 - 3m^2 + n}{(2n+3)(2n-1)}, \\ c_{nm} &= \frac{(n+1)(n+m)(n+m-1)}{(2n+1)(2n-1)}. \end{aligned} \quad (39-47)$$

We substitute (39-43) into (39-40) and take (39-46) into account. Furthermore we note that, e.g.,

$$\sum_{n=2}^{\infty} c_{nm} A_{nm} P_{n-2,m} = \sum_{n=0}^{\infty} c_{n+2,m} A_{n+2,m} P_{nm} . \quad (39-48)$$

In this way we can transform (39-40) into ellipsoidal coordinates β, λ , obtaining

$$T = \sum \sum P_{nm}(\sin \beta) (C_{nm} \cos m \lambda + D_{nm} \sin m \lambda) , \quad (39-49)$$

where

$$\begin{aligned} C_{nm} &= A_{nm} + \frac{1}{2} e^2 K_{nm} , & D_{nm} &= B_{nm} + \frac{1}{2} e^2 L_{nm} , \\ K_{nm} &= a_{n-2,m} A_{n-2,m} + b_{nm} A_{nm} + c_{n+2,m} A_{n+2,m} , \\ L_{nm} &= a_{n-2,m} B_{n-2,m} + b_{nm} B_{nm} + c_{n+2,m} B_{n+2,m} , \end{aligned} \quad (39-50)$$

the coefficients a_{nm} etc. being given by (39-47).

Let us abbreviate (39-49) as

$$T = \sum \sum T_{nm}(\beta, \lambda) . \quad (39-51)$$

This gives the potential T at the surface of the ellipsoid. In the space outside the ellipsoid, T can then be expressed as

$$T(u, \beta, \lambda) = \sum \sum S_{nm}(u) T_{nm}(\beta, \lambda) , \quad (39-52)$$

where

$$S_{nm}(u) = \frac{Q_{nm}(i \frac{u}{E})}{Q_{nm}(i \frac{b}{E})} , \quad (39-53)$$

Q_{nm} being a Legendre function of the second kind and u denoting the semi-minor axis of the coordinate ellipsoid passing through the space point under consideration; u, β, λ are spatial ellipsoidal coordinates. Cf. (Heiskanen and Moritz, secs. 1-19 and 1-20).

Now we can apply the basic boundary condition (2-32), taken on the ellipsoid:

$$\Delta g = - \frac{\partial T}{\partial h} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} T . \quad (39-54)$$

This relation expresses Δg in terms of T and of its derivative $\partial T / \partial h$ normal to the ellipsoid. This derivative equals

$$\frac{\partial}{\partial h} = \frac{\partial}{\partial s_u} = \frac{1}{w_0} \frac{\partial}{\partial u} \quad (39-55)$$

where

$$w_0 = \frac{1}{a} \sqrt{a^2 \sin^2 \beta + b^2 \cos^2 \beta} ; \quad (39-56)$$

cf. (Heiskanen and Moritz, 1967, pp.68-69).

Substituting (39-52) into (39-54) we obtain

$$\Delta g = \sum \sum \left(- \frac{\partial S_{nm}}{\partial h} + \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} S_{nm} \right)_0 T_{nm}(\beta, \lambda) , \quad (39-57)$$

where the symbol $()_0$ signifies that the quantity within parentheses is taken at the ellipsoid, that is, for $u = b$.

From (39-53) we get

$$\frac{dS_{nm}}{du} = \frac{i}{E} \frac{Q'_{nm}(z)}{Q_{nm}(z_0)} , \quad (39-58)$$

where

$$z = i \frac{u}{E} , \quad z_0 = i \frac{b}{E} , \quad (39-59)$$

$$Q'_{nm} = \frac{dQ_{nm}}{dz} . \quad (39-60)$$

It is now appropriate to use the series expansion (Hobson, 1931, p.195, eq.(19)):

$$Q_{nm}(z) = C(1-z^2)^{\frac{m}{2}} \left(\frac{1}{z^{n+m+1}} + \frac{(n+m+1)(n+m+2)}{2(2n+3)z^{n+m+3}} + \dots \right), \quad (39-61)$$

where C denotes a constant whose value is not needed for the present purpose.

We differentiate (39-61) with respect to z , substitute in (39-58) and put $z = z_0$ (that is, $u = b$), obtaining

$$\left(\frac{dS_{nm}}{du} \right)_0 = \frac{i}{E} \frac{z_0}{1-z_0^2} (n+1) \left(1 - \frac{(n+m+1)(n-m+1)}{(n+1)(2n+3)} z_0^{-2} + \dots \right). \quad (39-62)$$

Using the second equation in (39-59), it is straightforward to compute that

$$\frac{i}{E} \frac{z_0}{1-z_0^2} = -\frac{b}{a^2}; \quad (39-63)$$

furthermore, by (2-15),

$$-z_0^{-2} = e^{i2} = e^2 + e^4 + e^6 + \dots \quad (39-64)$$

differs from e^2 only by higher-order terms to be neglected. Thus (39-62) reduces to

$$\left(\frac{dS_{nm}}{du} \right)_0 = -\frac{b}{a^2} \left(n+1 + \frac{(n+m+1)(n-m+1)}{2n+3} e^2 \right). \quad (39-65)$$

By (39-55) we have

$$\left(\frac{\partial S_{nm}}{\partial h} \right)_0 = \frac{1}{w_0} \left(\frac{dS_{nm}}{du} \right)_0. \quad (39-66)$$

The expansion of (39-56) gives

$$\frac{1}{w_0} = 1 + \frac{1}{2} e^2 \cos^2 \beta, \quad (39-67)$$

neglecting e^4 and higher powers, which we are doing consistently all the time. This provides the first summand between parentheses in (39-57).

To get the second summand, we note that, by (39-53),

$$(S_{nm})_0 = S_{nm}(b) = 1. \quad (39-68)$$

By (Heiskanen and Moritz, 1967, p.78) we have

$$\left(\frac{1}{Y} \frac{\partial Y}{\partial h}\right)_0 = -\frac{2}{a} (1+f+m-2f \sin^2 \phi), \quad (39-69)$$

whence, by (39-11) and (39-12),

$$\left(\frac{1}{Y} \frac{\partial Y}{\partial h} S_{nm}\right)_0 = -\frac{2}{a} (1+e^2-e^2 \sin^2 \phi). \quad (39-70)$$

In terms of the mean radius (39-30) there is

$$\begin{aligned} \frac{1}{a} &= \frac{1}{R} \left(1 - \frac{1}{6} e^2\right), \\ \frac{b}{a^2} &= \frac{1}{R} \left(1 - \frac{2}{3} e^2\right). \end{aligned} \quad (39-71)$$

In view of eqs. (39-65) through (39-71), eq. (39-57) takes the form

$$\begin{aligned} \Delta g &= \frac{1}{R} \sum \sum \left[n - 1 + e^2 \left(\frac{(n+m+1)(n-m+1)}{2n+3} - \frac{n+1}{6} - \frac{n-3}{2} \sin^2 \phi \right) \right] \\ &\quad \cdot T_{nm}(\beta, \lambda). \end{aligned} \quad (39-72)$$

There remains to express $T_{nm}(\beta, \lambda)$ in terms of (ϕ, λ) . Remember that, by (39-49) and (39-51),

$$T_{nm}(\beta, \lambda) = P_{nm}(\sin \beta) (C_{nm} \cos m \lambda + D_{nm} \sin m \lambda). \quad (39-73)$$

In (39-43) we may replace β by ϕ in the term multiplied by e^2 , without impairing the accuracy, whence

$$P_{nm}(\sin \beta) = P_{nm}(\sin \phi) - \frac{1}{2} e^2 \frac{dP_{nm}}{d\phi} \sin \phi \cos \phi. \quad (39-74)$$

In (39-73) we now substitute $P_{nm}(\sin\theta)$ by (39-74) and the coefficients C_{nm} and D_{nm} by the first line of (39-50). The expression for $T_{nm}(\theta, \lambda)$ thus obtained is then inserted into (39-72). After some straightforward algebra, neglecting again terms multiplied by e^4 , we get

$$\begin{aligned} \Delta g = & \sum \sum \frac{n-1}{R} P_{nm}(\sin\phi) (A_{nm} \cos m\lambda + B_{nm} \sin m\lambda) + \\ & + \frac{e^2}{R} \sum \sum \left\{ \left[\left(\frac{(n+m+1)(n-m+1)}{2n+3} - \frac{n+1}{6} \right) P_{nm}(\sin\phi) - \right. \right. \\ & - \frac{n-3}{2} \sin^2\phi P_{nm}(\sin\phi) - \frac{n-1}{2} \sin\phi \cos\phi \frac{dP_{nm}}{d\phi} \Big] \cdot \\ & \cdot (A_{nm} \cos m\lambda + B_{nm} \sin m\lambda) + \\ & \left. \left. + \frac{n-1}{2} P_{nm}(\sin\phi) (K_{nm} \cos m\lambda + L_{nm} \sin m\lambda) \right\} . \end{aligned} \quad (39-75)$$

We further express $\sin\phi \cos\phi dP_{nm}/d\phi$ by (39-46), with ϕ instead of θ , and $\sin^2\phi P_{nm}$ by

$$\sin^2\phi P_{nm}(\sin\phi) = \alpha_{nm} P_{n+2,m} + \beta_{nm} P_{nm} + \gamma_{nm} P_{n-2,m} , \quad (39-76)$$

where

$$\begin{aligned} \alpha_{nm} &= \frac{(n-m+1)(n-m+2)}{(2n+1)(2n+3)} , & \beta_{nm} &= \frac{2n^2-2m^2+2n-1}{(2n+3)(2n-1)} , \\ \gamma_{nm} &= \frac{(n+m)(n+m-1)}{(2n+1)(2n-1)} ; \end{aligned} \quad (39-77)$$

this relation is obtained by applying eq. (39-45) twice. At last we change the summation variable in terms containing $P_{n+2,m}$ and $P_{n-2,m}$ similarly to (39-48).

The final result is

$$\Delta g = \Delta g^0 + e^2 \Delta g^1 \quad (39-78)$$

where

$$\Delta g^0 = \sum_{n=2}^{\infty} \sum_{m=0}^n \frac{n-1}{R} P_{nm}(\sin\phi) (A_{nm} \cos m\lambda + B_{nm} \sin m\lambda) , \quad (39-79)$$

$$\Delta g^1 = \frac{1}{R} \sum_{n=2}^{\infty} \sum_{m=0}^n P_{nm}(\sin \phi) (G_{nm} \cos m \lambda + H_{nm} \sin m \lambda) \quad (39-80)$$

with

$$G_{nm} = \kappa_{nm} A_{n-2,m} + \lambda_{nm} A_{nm} + \nu_{nm} A_{n+2,m} \quad (39-81)$$

$$H_{nm} = \kappa_{nm} B_{n-2,m} + \lambda_{nm} B_{nm} + \nu_{nm} B_{n+2,m}$$

and

$$\begin{aligned} \kappa_{nm} &= - \frac{3(n-3)(n-m-1)(n-m)}{2(2n-3)(2n-1)}, \\ \lambda_{nm} &= \frac{n^3 - 3m^2n - 9n^2 - 6m^2 - 10n + 9}{3(2n+3)(2n-1)}, \\ \nu_{nm} &= - \frac{(3n+5)(n+m+2)(n+m+1)}{2(2n+5)(2n+3)}. \end{aligned} \quad (39-82)$$

The computation of the correction term (39-80), which is multiplied by e^2 and has, therefore, a very small effect $e^2 \Delta g^1$, can be done in an approximate way, using a truncated spherical-harmonic expansion obtained by a combination of satellite and gravimetric data such as given in (Lerch et al., 1977), (Rapp, 1977) or (Gaposchkin, 1979)¹.

The comparison between (39-40) and (39-79) shows that $T = T^0$ and Δg^0 is connected by the usual spherical relation; cf. (Heiskanen and Moritz, 1967, p.89, eq.(2-155')); this is as it should be.

Collocation with ellipsoidal corrections. If we wish to apply ellipsoidal corrections in least-squares collocation we may proceed in the following manner.

The analytical structure underlying our whole preceding treatment of least-squares collocation is based on a spherical reference surface. In fact, as we have seen repeatedly, this analytical structure of the anomalous gravitational field is represented by covariance propagation on the basis of spherical formulas such as (15-7) to (15-9); also the covariance function $K(P,Q)$, being homogeneous and isotropic, is essentially related to the sphere.

¹ The summation in (39-80) starts with $n=2$, although zero and first degree terms are introduced by (39-48); these latter terms, however, are conventionally suppressed; cf. (Heiskanen and Moritz, 1967, p.97).

Thus least-squares collocation formulas such as (11-27) (using s instead of ξ),

$$s = C_{s1} C_{11}^{-1} l, \quad (39-83)$$

should be written, in the present notation, as

$$s^0 = C_{s1} C_{11}^{-1} l^0, \quad (39-84)$$

putting

$$s = s^0 + e^2 s^1, \quad l = l^0 + e^2 l^1 \quad (39-85)$$

in conformity with (39-3). Combining (39-84) and (39-85) gives

$$s = C_{s1} C_{11}^{-1} (l - e^2 l^1) + e^2 s^1, \quad (39-86)$$

which is the improvement of (39-83) by applying ellipsoidal corrections.

The procedure expressed by (39-86) may be described as follows:

1. reduction of the data l from the ellipsoid to the sphere by subtracting $e^2 l^1$;
2. application of the spherical collocation formula (39-84), giving $C_{s1} C_{11}^{-1} (l - e^2 l^1)$;
3. reduction of this result from the sphere back to the ellipsoid by adding $e^2 s^1$.

It is thus quite similar to reduction procedures described at the end of the preceding section. The ellipsoidal reductions $e^2 l^1$ and $e^2 s^1$ are given by the basic reduction formulas derived in this section: (39-21) for the geoidal height N , (39-38) and (39-39) for the components ξ and η of the deflection of the vertical, and (39-80) for the gravity anomaly Δg ; there is, by definition, no ellipsoidal reduction for the anomalous potential T , cf. (39-6). We recall that for all ellipsoidal corrections we may use an approximate anomalous field as defined, e.g., by one of the known truncated spherical-harmonic expansions.

PART D

THE GEODETIC BOUNDARY-VALUE PROBLEM

The geodetic boundary-value problem is the determination of the earth's physical surface from the values of the gravity vector and the gravity potential given on it. The book treats recent developments in this field, restricting itself, however, to two subjects: first, the investigation of series solutions by Molodensky, Brovar, and others, of their convergence, and of their equivalence (secs.43-49); and secondly, mathematical studies on the existence and uniqueness of the solution performed recently by Hörmander, Krarup, and Sansø (secs.50-54).

Secs. 40-42 provide an introduction to the mathematical structure of Molodensky's problem; a previous knowledge of the basic principles, such as given in chapter 8 of (Heiskanen and Moritz, 1967), is desirable though not indispensable.

The reader primarily interested in applications will find information on computational formulas and their theoretical background in secs. 40-49.

Hörmander's work on existence and uniqueness represents a mathematical tour de force of great depth and complexity. The present treatment attempts only a general, non-technical description of the method, the mathematical background, and the results.

Sansø's treatment of the geodetic boundary-value problem provides a highly original and simple new approach: by means of a Legendre transformation, the free boundary-value problem is transformed into a problem with a fixed boundary. This approach is related to Molodensky's theory in much the same way as Hamilton's treatment of classical mechanics is related to the Newtonian approach; it is of similar beauty. The book treats rather fully the elementary aspects of Sansø's theory.

Geodynamical effects such as time variation of reference systems and tidal influences are small but of great principal significance. Therefore, an outline of them concludes the book.

40. MOLODENSKY'S PROBLEM

In sec. 2 we have briefly mentioned the *problem of Stokes*, the gravimetric determination of the geoid; the solution is given by Stokes' formula (2-35). This approach, though mathematically quite simple, meets with conceptual difficulties because it presupposes that all measurements refer to the geoid. Now the actual geodetic measurements are made at the physical earth's surface, or topographic surface, which is the earth's surface which we see and on which we walk. Therefore, these measurements must be reduced to sea level, that is, to the geoid. This, however, implies the knowledge of the density of the masses above the geoid, and this density is only imperfectly accessible to observation. The practical impact of this difficulty is not forbidding, but the conceptual difficulties remain.

Therefore, M.S. Molodensky in 1945 proposed a different approach, namely the direct gravimetric determination of the physical earth's surface. This *problem of Molodensky* has played a fundamental role in theoretical geodesy during the last decades. A review of the results obtained by 1960 is given in the monograph (Molodenskii et al., 1962). An elementary presentation with emphasis on the physical background is found in chapter 8 of (Heiskanen and Moritz, 1967). Our present treatment will stress the mathematical structure of Molodensky's problem and present some recent developments.

The problems of both Stokes and Molodensky deal with data given on a surface (the geoid and the physical earth's surface, respectively), which is a boundary for the region outside this surface. They are thus *geodetic boundary-value problems*.

The problem of Molodensky may be formulated briefly as follows: given, at all points of the physical earth's surface S , the gravity potential W and the gravity vector \underline{g} , to determine the surface S . The potential W can be determined by leveling combined with gravity measurements; this gives the potential up to an unknown constant which, however, can be found indirectly by other methods, especially distance measurements. The magnitude of the gravity vector \underline{g} , which is gravity g , is measured by gravimetry, and the direction of \underline{g} , which is the plumb line, is obtained by astronomical measurements of latitude ϕ and longitude λ . It is assumed that these measurements have been corrected for luni-solar tidal effects and other temporal variations, so that our problem is independent of time. We further suppose that the very small effect of the atmosphere has been taken into account by an appropriate reduction. Hence, the space outside the surface S can be considered empty.

We thus assume that the earth is a rigid body which rotates with a constant and known angular velocity ω around a fixed axis, which passes

through the earth's center of mass. This center of mass will be taken as the origin 0 of a cartesian coordinate system, the x_1 -axis coinciding with the axis of rotation.

The gravitational potential V is a harmonic function outside S . For large values of the radius vector

$$r = ||\underline{x}|| = \sqrt{x_1^2 + x_2^2 + x_3^2} \quad (40-1)$$

it has an expansion in spherical harmonics of the form (6-2):

$$V(\underline{x}) = \frac{GM}{r} + \frac{Y_1(\theta, \lambda)}{r^2} + \frac{Y_2(\theta, \lambda)}{r^3} + \dots, \quad (40-2)$$

where G is the gravitational constant, M denotes the total mass of the earth, and $Y_n(\theta, \lambda)$ are Laplace surface harmonics (3-22), θ (polar distance) and λ (longitude) forming together with the radius vector r a system of spherical coordinates related to the cartesian coordinates $\underline{x} = [x_1, x_2, x_3]$ in the usual way:

$$\begin{aligned} x_1 &= r \sin \theta \cos \lambda, \\ x_2 &= r \sin \theta \sin \lambda, \\ x_3 &= r \cos \theta. \end{aligned} \quad (40-3)$$

The condition that the coordinate origin 0 coincides with the center of mass implies that the spherical harmonics of first degree vanish identically:

$$Y_1(\theta, \lambda) \equiv 0, \quad (40-4)$$

so that V must have the form

$$V(\underline{x}) = \frac{GM}{r} + O\left(\frac{1}{r^3}\right) \quad \text{for} \quad r \rightarrow \infty. \quad (40-5)$$

The gravity potential W is then given by

$$W(\underline{x}) = V(\underline{x}) + \frac{1}{2} \omega^2 (x_1^2 + x_2^2). \quad (40-6)$$

It will also be assumed that the surface S is a closed, simply connected and smooth surface, being differentiable as often as required.

It may be questioned whether Molodensky's problem thus formulated is today geodetically relevant at all. On the one hand, the prerequisites for Molodensky's problem, especially continuous coverage of the whole earth's surface by gravity measurements, are still far from being realized; on the other hand, there are many more data of different kind, such as satellite data, that transcend the frame of Molodensky's problem and must be handled by data combination techniques such as least-squares collocation.

To these questions we may answer as follows. From a practical point of view, the integral formulas arising in the solution of boundary-value problems are often computationally more convenient than collocation and retain their importance if gravity data are available to a sufficient extent, at least locally (cf. sec.49). From a theoretical point of view, the geodetic boundary-value problem represents an especially interesting and significant special case, whose importance for the conceptual structure of geodesy, from the time of Clairaut¹ to the present day, can hardly be overestimated; interestingly enough, the theory was always far ahead of the data available at the time. In fact, the consecutive stages in the development of the boundary-value problem--Clairaut, Stokes, Molodensky--always served as measures of perfection for geodetic theory and set new standards.

Even today Molodensky's problem is not yet completely clarified from a mathematical point of view, with respect to existence and uniqueness of the solution, in spite of the decisive progress made in the last few years; it remains a challenge to theoreticians.

Let us now try to get a first grasp of the mathematical nature of Molodensky's problem.

The gravity vector \underline{g} can be expressed in terms of measured gravity g and of astronomical latitude ϕ and longitude Λ as

$$\underline{g} = \begin{bmatrix} g \cos \phi \cos \Lambda \\ g \cos \phi \sin \Lambda \\ g \sin \phi \end{bmatrix} . \quad (40-7)$$

In space the vector \underline{g} and the potential W may be considered functions of the rectangular coordinates:

$$\underline{g} = \underline{g}(x_1, x_2, x_3) , \quad W = W(x_1, x_2, x_3) . \quad (40-8)$$

¹The well-known formula of Clairaut (cf. Heiskanen and Moritz, 1967, p.69) provides a relation between gravity and the geometric form of a level surface. In this sense, Clairaut is a predecessor of Stokes and Molodensky.

On the earth's surface S , they are functions of two surface coordinates, for which we may take the astronomical coordinates ϕ and λ :

$$\bar{\underline{g}} = \bar{\underline{g}}(\phi, \lambda), \quad \bar{W} = \bar{W}(\phi, \lambda); \quad (40-9)$$

the overbar denotes restriction of spatial functions to the surface S , whereas underlining characterizes vectors and matrices as usual.

Now $\bar{\underline{g}}$ may be expressed, in a certain sense, as a function of S and \bar{W} , symbolically

$$\bar{\underline{g}} = F(S, \bar{W}). \quad (40-10)$$

This means that, given the surface S and the gravity potential \bar{W} on it, the gravity vector $\bar{\underline{g}}$ on S is then uniquely determined and can be computed.

In fact, this may be done as follows. Let S and \bar{W} be given. Compute the centrifugal potential on S (which can be done since the surface S is supposed to be given and consequently the coordinates x_1, x_2, x_3 of the surface points are known) and subtract it from \bar{W} ; this gives the gravitational potential \bar{V} on S . From \bar{V} on S we get the potential V outside S by solving Dirichlet's boundary value problem, which has a unique solution. Now

$$\underline{g} = \text{grad } V + \text{centrifugal force}$$

(grad denoting the gradient) can be computed outside S and, by the continuity of first derivatives, also on S , giving $\bar{\underline{g}}$. Thus $\bar{\underline{g}}$ is, in fact, uniquely determined by S and \bar{W} , so that (40-10) holds. In the terminology of sec. 5, the function F is a nonlinear operator or mapping.

Suppose now that it were possible to solve (40-10) for S :

$$S = \phi(\bar{W}, \bar{\underline{g}}). \quad (40-11)$$

This would express the earth's surface S in terms of \bar{W} and $\bar{\underline{g}}$, solving Molodensky's problem.

This is probably the conceptually simplest formulation of Molodensky's problem. However, the transition from (40-10) to (40-11) is mathematically extremely difficult. If S , \bar{W} and $\bar{\underline{g}}$ were simple real numbers and F were an ordinary function (supposed sufficiently smooth) of two real variables, then the solution of (40-10) for S would be straightforward. The

existence of such a solution is guaranteed by the elementary implicit function theorem.

In reality the function F in (40-10) is a rather complicated nonlinear operator, and the existence of a solution (40-11) is by no means obvious. There are implicit function theorems for nonlinear operators (e.g. Dieudonné, 1960; Loomis and Sternberg, 1968; Schwartz, 1969; Sternberg, 1969), but the conditions for their application are not satisfied in the geodetic case. It was the merit of Hörmander (1976) to have found, by a mathematical *tour de force*, an implicit function theorem that is applicable to the geodetic boundary-value problem (sec.51).

To get some first insight into the matter, let us forget all mathematical difficulties and proceed formally as if S , \bar{W} , and \bar{g} were simply real numbers and F were a simple function. Since \bar{W} is given, it can be considered fixed once and for all, so that (40-10) becomes a function of S only:

$$\bar{g} = f(S) . \quad (40-12)$$

To further simplify the notation, we write g instead of \bar{g} , obtaining

$$g = f(S) . \quad (40-13)$$

Thus S is simply given by the inverse function

$$S = f^{-1}(g) , \quad (40-14)$$

so that the implicit function problem reduces to an inverse function problem.

To practically find this inverse function, that is, to solve (40-13) for S , we may apply the usual procedure for solving nonlinear equations, namely *linearization*.

Let us introduce an approximation S_0 to the earth's surface S and let g_0 be the corresponding gravity vector, related to S_0 by (40-13):

$$g_0 = f(S_0) . \quad (40-15)$$

Write, formally,

$$\begin{aligned} S &= S_0 + \Delta S , \\ g &= g_0 + \Delta g \end{aligned} \quad (40-16)$$

and apply Taylor's theorem to (40-13):

$$g_0 + \Delta g = f(S_0 + \Delta S) = f(S_0) + f'(S_0)\Delta S ,$$

omitting quadratic and higher terms. In view of (40-15) this becomes

$$\Delta g = f'(S_0)\Delta S . \quad (40-17)$$

The formal solution of this equation is

$$\Delta S = [f'(S_0)]^{-1}\Delta g . \quad (40-18)$$

Let us link these ideas with the conventional approach to Molodensky's problem. Here S_0 is the telluroid and g_0 is normal gravity on it; Δg is the usual gravity anomaly referred to the earth's surface (it is here possible to disregard the original vector character of Δg and regard it as a scalar quantity) and ΔS is represented by the height anomaly ζ characterizing the separation between earth's surface S and telluroid S_0 ; for details cf. secs. 41 and 42. Thus (40-18) becomes

$$\zeta = M\Delta g , \quad (40-19)$$

where $M = [f'(S_0)]^{-1}$ denotes the linear Molodensky operator computing ζ from Δg . Practically one uses Stokes' formula with suitable corrections, as the following sections will show.

Higher approximations may be obtained by *Newton's method*. Combining (40-15), (40-16) and (40-18) we get

$$S_1 = S_0 + [f'(S_0)]^{-1}[g - f(S_0)] , \quad (40-20)$$

where we have written S_1 instead of S to indicate that by this equation we get a better approximation S_1 rather than the true value S itself. By repeated application of this formula we get successive better approximations S_2, S_3, \dots

$$\begin{aligned} S_2 &= S_1 + [f'(S_1)]^{-1}[g - f(S_1)] , \\ S_3 &= S_2 + [f'(S_2)]^{-1}[g - f(S_2)] , \\ &\vdots \end{aligned} \quad (40-21)$$

Graphically Newton's procedure is illustrated by Fig. 40.1. The unknown abscissa S for the given ordinate g is approached by following the broken line with arrows.

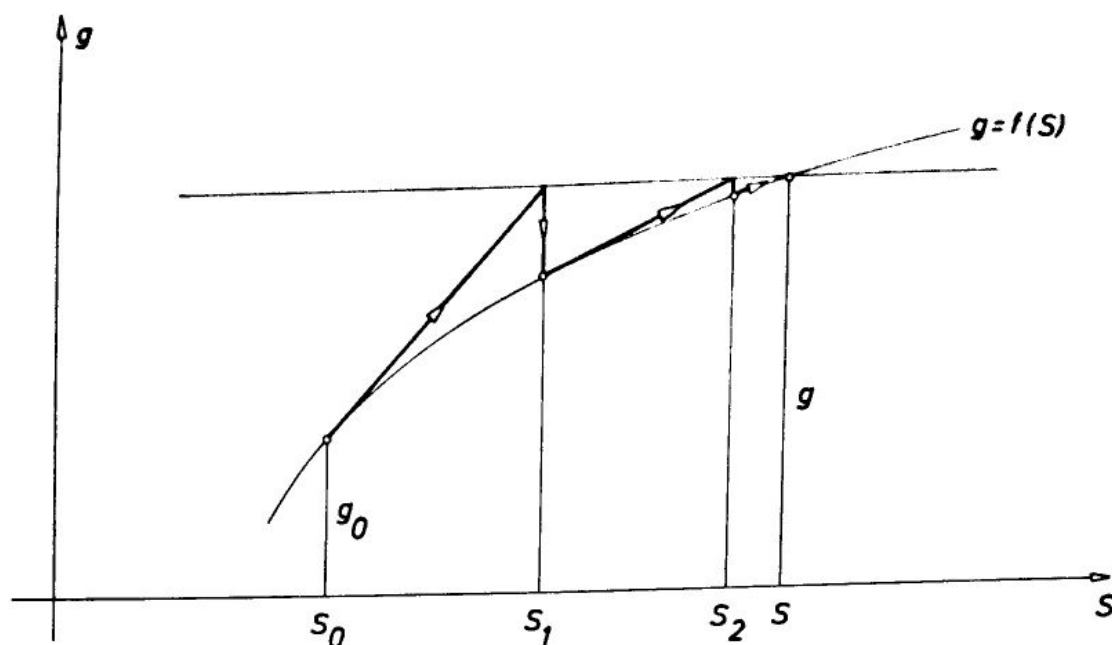


FIGURE 40.1. *Newton's method.*

The convergence of Newton's procedure is known to be very good, namely quadratic: there is a constant K independent of n such that

$$|S_{n+1} - S_n| \leq K |S_n - S_{n-1}|^2. \quad (40-22)$$

The linearized problem is important both in its own right and as a step in the solution of the nonlinear problem; therefore, the next sections will be devoted to it.

41. LINEARIZATION

The conventional linearizations of Molodensky's problem as given, e.g., in (Molodenskii et al., 1962, chapter V) or in (Heiskanen and Moritz, 1967, sec. 8-5) are practically sufficiently accurate but not completely rigorous.

Rigorous linearizations have been given by Meissl (1971b) and Krarup (1973); we shall follow the latter.

As usual, the linearization consists in introducing suitable known approximate values and applying Taylor's theorem. Thus the physical earth's surface S will be approximated by a known surface, close to S , which will be called the *telluroid* and denoted by Σ . The points Q of Σ are thought to be in some one-to-one correspondence with the points P of S ; cf. Fig. 41.1. We also introduce a normal potential U which constitutes an analytical approximation to the actual gravity potential W ; U is usually taken as the gravity potential of an equipotential ellipsoid.

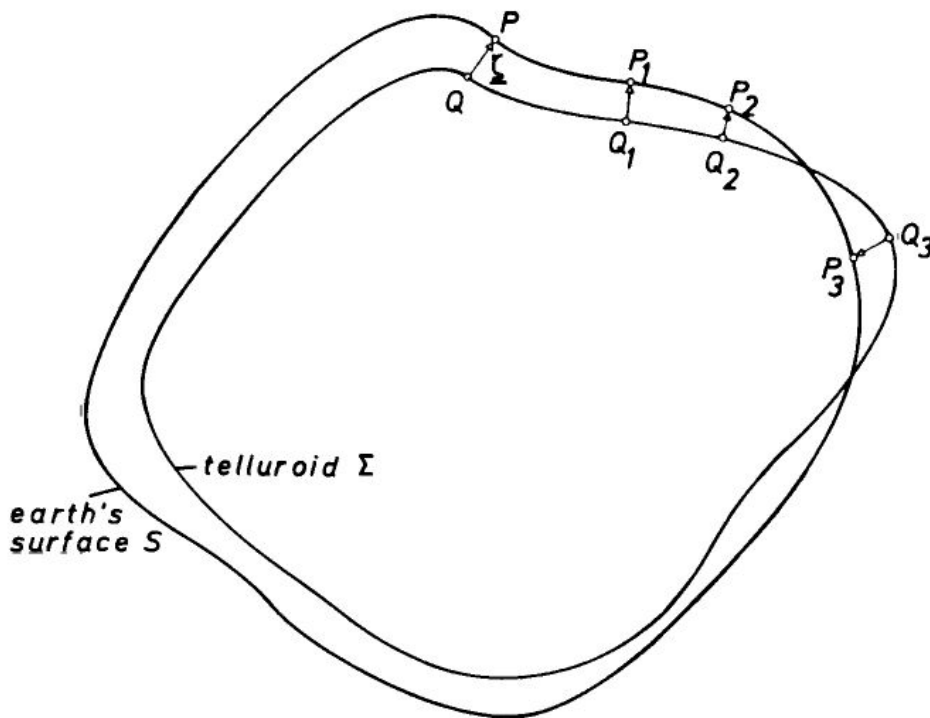


FIGURE 41.1. The telluroid Σ as an approximation to the earth's surface S .

Let

$$\underline{y} = \text{grad } U \quad (41-1)$$

denote the normal gravity vector, in the same way as

$$\underline{g} = \text{grad } W \quad (41-2)$$

expresses the actual gravity vector.

Since Σ and U are given, we can compute U and \underline{y} at Q , that is, U_Q and \underline{y}_Q . As potential W and gravity \underline{g} are supposed to be given on S (in the notation of sec.40, they are \bar{W} and $\bar{\underline{g}}$), we know them at every point P on S , that is, we know W_P and \underline{g}_P . We, therefore, can compute the differences

$$\Delta W = W_P - U_Q, \quad (41-3)$$

$$\Delta \underline{g} = \underline{g}_P - \underline{y}_Q, \quad (41-4)$$

called *potential anomaly* and (vectorial) *gravity anomaly*, respectively.

By appropriate definitions of the telluroid it is possible to make one of the two quantities (41-3) and (41-4) equal to zero. To have

$$\Delta W = 0 \quad (41-5)$$

(this means zero potential anomaly, not Laplace's equation!) we may define Q by the three conditions

$$U_Q = W_P, \quad \phi_Q = \phi_P, \quad \lambda_Q = \lambda_P. \quad (41-6)$$

Here ϕ and λ are given by

$$\underline{y} = \begin{bmatrix} \gamma \cos \phi \cos \lambda \\ \gamma \cos \phi \sin \lambda \\ \gamma \sin \phi \end{bmatrix}. \quad (41-7)$$

in complete analogy to (40-7); thus the normal latitude ϕ and longitude λ determine the direction of the normal gravity vector \underline{y} , in the same way as ϕ and λ define the direction of \underline{g} . The surface formed by the points Q in this manner has been called by Krarup (1973) the *Marussi telluroid* because the three "Marussi coordinates" potential, latitude and longitude are identified.

In this way, the potential anomaly ΔW can be made zero. Somewhat surprising at first sight is that also the gravity anomaly $\Delta \underline{g}$ can be made to vanish. This requires defining the points Q of the telluroid by

$$\underline{y}_Q = \underline{g}_P. \quad (41-8)$$

Expressing this vector condition in terms of magnitude and direction of the vectors involved, we get three conditions

$$\begin{aligned} \gamma_Q &= g_P, \\ \phi_Q &= \phi_P, \\ \lambda_Q &= \lambda_P, \end{aligned} \tag{41-9}$$

which again completely determine Q . Since g, ϕ, λ may be called "gravimetric coordinates", the corresponding locus of points Q has been called by Krarup the *gravimetric telluroid*; for it, in fact,

$$\Delta g = 0. \tag{41-10}$$

After these possible specializations, let us return to the general case in which both ΔW and Δg are nonzero. As usual, we define the disturbing potential T by

$$T = W - U, \tag{41-11}$$

W and U referring to the same point (this distinguishes T from the potential anomaly ΔW , in which W and U refer to different points!).

On substituting

$$W_P = U_P + T_P \tag{41-12}$$

we get from (41-3) and (41-4)

$$T_P + U_P - U_Q = \Delta W, \tag{41-13}$$

$$g_P - \gamma_Q = \Delta g. \tag{41-14}$$

Let us now proceed with the linearization. We put

$$\underline{\zeta} = \text{vector } QP \tag{41-15}$$

(see Fig.41.1) and systematically neglect all quantities of second and higher order in $\underline{\zeta}$. It is well known and easy to see that quantities such as T and Δg have the same order of magnitude as $\underline{\zeta}$. So also $T^2, T\underline{\zeta}$, etc., are quantities of second order to be neglected.

By a Taylor expansion restricted to linear terms we get

$$U_P = U_Q + \text{grad } U \cdot \underline{\xi} = U_Q + \underline{\gamma} \cdot \underline{\xi} \quad (41-16)$$

where the dot denotes the inner product of two vectors. Let us proceed in the same way with the normal gravity vector:

$$\underline{\gamma}_P = \underline{\gamma}_Q + \text{grad } \underline{\gamma} \cdot \underline{\xi} \quad (41-17)$$

What is $\text{grad } \underline{\gamma}$? To see this, let us write this equation in index notation, using the summation convention (summation over an index that occurs twice in a product, in our case over j):

$$\gamma_{P,i} = \gamma_{Q,i} + \frac{\partial \gamma_i}{\partial x_j} \xi_j = \gamma_{Q,i} + M_{ij} \xi_j \quad (41-18)$$

where

$$M_{ij} = \frac{\partial \gamma_i}{\partial x_j} = \frac{\partial}{\partial x_j} \left(\frac{\partial U}{\partial x_i} \right) = \frac{\partial^2 U}{\partial x_i \partial x_j} \quad (41-19)$$

Hence $\text{grad } \underline{\gamma}$ is nothing else than the matrix

$$\underline{M} = [M_{ij}] = \left[\frac{\partial^2 U}{\partial x_i \partial x_j} \right] \quad (41-20)$$

formed by the second derivatives of the normal potential U . Therefore, we may write (41-17) as

$$\underline{\gamma}_Q = \underline{\gamma}_P - \underline{M} \underline{\xi} \quad (41-21)$$

It is clear that $\underline{\gamma}$ in (41-16) and \underline{M} in (41-21) refer to point Q .

Let us similarly expand T_P :

$$T_P = T_Q + \text{grad } T \cdot \underline{\xi} \quad (41-22)$$

Now, however, $\text{grad } T$ is already small of first order, so that $\text{grad } \underline{T} \cdot \underline{\xi}$ is of second order and, therefore, negligible. Thus, consistent with our linear approximation, we simply have

$$T_P = T_Q, \quad (41-22)$$

The insertion of (41-16), (41-21), and (41-22) into (41-13) and (41-14) now gives

$$T_Q + \underline{Y} \cdot \underline{\zeta} = \Delta W, \quad (41-23)$$

$$\underline{g}_P - \underline{Y}_P + \underline{M}\underline{\zeta} = \Delta \underline{g}. \quad (41-24)$$

Furthermore,

$$\begin{aligned} \underline{g}_P - \underline{Y}_P &= (\text{grad } W)_P - (\text{grad } U)_P \\ &= \text{grad } (W - U)_P \\ &= (\text{grad } T)_P \\ &\doteq (\text{grad } T)_Q, \end{aligned}$$

for the same reason as (41-22). We thus finally get

$$T + \underline{Y}^T \underline{\zeta} = \Delta W, \quad (41-25)$$

$$\text{grad } T + \underline{M}\underline{\zeta} = \Delta \underline{g}, \quad (41-26)$$

in which T and $\text{grad } T$ refer to Q , as well as \underline{Y} and \underline{M} . We have used the matrix notation $\underline{a}^T \underline{b}$ for the inner product $\underline{a} \cdot \underline{b}$, the transpose of \underline{a} being denoted by \underline{a}^T . The reader will note analogies between the present section and sec. 27; cf. p. 234.

These two equations will be basic for our further developments. Let us solve (41-26) for $\underline{\zeta}$, assuming the matrix \underline{M} invertible,

$$\underline{\zeta} = \underline{M}^{-1}(\Delta \underline{g} - \text{grad } T), \quad (41-27)$$

and substitute into (41-25):

$$T + \underline{Y}^T \underline{M}^{-1}(\Delta \underline{g} - \text{grad } T) = \Delta W$$

or

$$T - \underline{Y}^T \underline{M}^{-1} \text{grad } T = \Delta W - \underline{Y}^T \underline{M}^{-1} \Delta \underline{g}. \quad (41-28)$$

On putting

$$\underline{m} = -\underline{M}^{-1} \underline{Y} \quad (41-29)$$

we get

$$\tau + \underline{m}^T \text{grad } \tau = \Delta W + \underline{m}^T \Delta \underline{g} \quad (41-30)$$

This equation, which holds on the telluroid Σ , constitutes the *fundamental boundary condition* for the linearized Molodensky problem. It is a generalization of the "fundamental equation of physical geodesy" from Stokes' to Molodensky's problem, just as (41-25) is a generalization of Bruns' formula (pp.14-15).

Various forms of the boundary condition. Let us introduce new coordinates q_i by

$$\begin{aligned} q_1 &= q_1(x_1, x_2, x_3) \ , \\ q_2 &= q_2(x_1, x_2, x_3) \ , \\ q_3 &= q_3(x_1, x_2, x_3) \ , \end{aligned} \quad (41-31)$$

or briefly

$$q_i = q_i(x_j) \ , \quad (41-32)$$

and let us assume that the inverse transformation

$$x_j = x_j(q_k) \quad (41-33)$$

also exists. More specifically, we shall select q_i to be the cartesian components of the normal gravity vector:

$$q_i = \gamma_i = \frac{\partial U}{\partial x_i} \quad (41-34)$$

It is clear that one-to-one relations (41-32) and (41-33) exist, at least in the spatial vicinity of the earth's surface, so that the quantities (41-34) may indeed be used as spatial curvilinear coordinates.

The matrix \underline{M} introduced by (41-19) and (41-20) may be written as

$$\underline{M} = \left[\frac{\partial y_i}{\partial x_j} \right] ; \quad (41-35)$$

it is, therefore, nothing else than the Jacobian matrix of the transformation (41-32). It is well known that the inverse matrix \underline{M}^{-1} is then simply the Jacobian matrix of the inverse transformation (41-33):

$$\underline{M}^{-1} = \left[\frac{\partial x_i}{\partial y_j} \right] . \quad (41-36)$$

This may also be shown directly: we have

$$\frac{\partial y_i}{\partial x_j} \frac{\partial x_j}{\partial y_k} = \frac{\partial y_i}{\partial y_k} \quad (41-37)$$

by the chain rule of differential calculus; furthermore

$$\frac{\partial y_i}{\partial y_k} = \delta_{ik} = \begin{cases} 1 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} ; \quad (41-38)$$

for instance, clearly

$$\frac{\partial y_1}{\partial y_1} = 1 , \quad \frac{\partial y_1}{\partial y_3} = 0 .$$

Therefore, (41-37) becomes

$$\frac{\partial y_i}{\partial x_j} \frac{\partial x_j}{\partial y_k} = \delta_{ik} , \quad (41-39)$$

which, by (41-35) and (41-36), is nothing but the equation

$$\underline{M} \underline{M}^{-1} = \underline{I} \quad (41-40)$$

in index notation, \underline{I} denoting the unit matrix.

Now the vector \underline{m} , defined by (41-29), becomes in index notation

$$m_i = - \frac{\partial x_i}{\partial \gamma_j} \gamma_j, \quad (41-41)$$

and we further have

$$\begin{aligned} \underline{m}^T \text{grad } T &= m_i \frac{\partial T}{\partial x_i} \\ &= - \frac{\partial T}{\partial x_i} \frac{\partial x_i}{\partial \gamma_j} \gamma_j \\ &= - \frac{\partial T}{\partial \gamma_j} \gamma_j, \end{aligned} \quad (41-42)$$

again by the chain rule. Hence (41-30) becomes

$$T - \gamma_i \frac{\partial T}{\partial \gamma_i} = f \quad (41-43)$$

where we have used the abbreviation

$$f = \Delta W + \underline{m}^T \Delta \underline{g}. \quad (41-44)$$

An even greater simplification is achieved by introducing "quasi-spherical coordinates" ρ, ϕ, λ by

$$\begin{aligned} \gamma_1 &= - \frac{1}{\rho^2} \cos \phi \cos \lambda, \\ \gamma_2 &= - \frac{1}{\rho^2} \cos \phi \sin \lambda, \\ \gamma_3 &= - \frac{1}{\rho^2} \sin \phi. \end{aligned} \quad (41-45)$$

Here ϕ and λ are normal latitude and longitude as before, because the vector γ_i is nothing else than normal gravity. The coordinate ρ is taken as positive. If the reference ellipsoid becomes a sphere, then ρ becomes proportional to the radius vector, as we shall see below, so that ρ, ϕ, λ become spherical coordinates; hence the name, quasi-spherical coordinates.

Now

$$\frac{\partial T}{\partial \rho} = \frac{\partial T}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \rho}, \quad (41-46)$$

again by the chain rule;

$$\frac{\partial \gamma_1}{\partial \rho} = \frac{2}{\rho^3} \cos \phi \cos \lambda = -\frac{2}{\rho} \gamma_1$$

and, generally,

$$\frac{\partial \gamma_i}{\partial \rho} = -\frac{2}{\rho} \gamma_i$$

by (41-45). Thus (41-46) becomes

$$\frac{\partial T}{\partial \rho} = -\frac{2}{\rho} \gamma_i \frac{\partial T}{\partial \gamma_i}, \quad (41-47)$$

and (41-43) reduces to

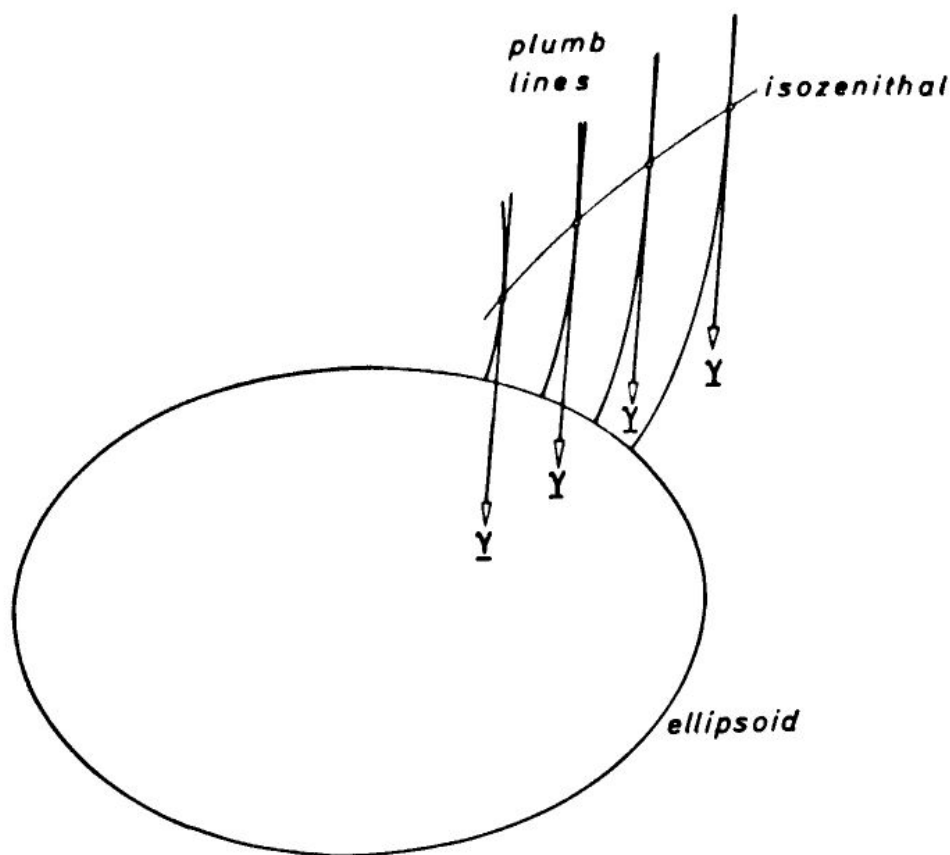
$$\rho \frac{\partial T}{\partial \rho} + 2T = 2f. \quad (41-48)$$

It should be pointed out that (41-48), in spite of its simplicity, is *rigorously* equivalent to (41-30); there is no further approximation involved.

What is the geometrical meaning of the derivative $\partial T / \partial \rho$? According to the definition of a partial derivative, $\partial / \partial \rho$ means differentiation with respect to one coordinate ρ , the two other coordinates ϕ, λ being held constant. This means differentiation along a line

$$\phi = \text{const.}, \quad \lambda = \text{const.} \quad (41-49)$$

Such lines are called *isozenithals* (with respect to the normal gravity field). The reason for this name is that (ϕ, λ) may be considered the coordinates of the (ellipsoidal) zenith on the celestial sphere. The isozenithals may also be regarded as the lines along which the normal gravity vectors are all parallel, having the same direction (41-49). If the plumb lines were straight lines, then the isozenithals would coincide with the plumb lines; as the normal plumb line curvature is quite small, isozenithals and plumb lines are not very different. For a detailed picture of the geometry of the normal gravity field cf. (Sünkel, 1978a); see also Fig.41.2.

FIGURE 41.2. *Plumb lines and an isozenithal.*

In view of the fundamental importance of our boundary condition, let us approach it from still another angle. Let τ denote the arc length of the isozenithal line, measured, e.g., from the ellipsoid positive upwards (so that it represents the height above the ellipsoid, measured along the isozenithal). Then $\partial/\partial\tau$ represents a derivative along the isozenithal, in the same way as $\partial/\partial\rho$. Therefore, these two derivatives, having the same direction, can only differ in scale, that is, they must be proportional:

$$\frac{\partial}{\partial\tau} = C \frac{\partial}{\partial\rho} , \quad (41-50)$$

To find the proportionality factor C we apply this equation to γ :

$$\frac{\partial\gamma}{\partial\tau} = C \frac{\partial\gamma}{\partial\rho} . \quad (41-51)$$

The right-hand side can be easily evaluated, since by (41-45)

$$\gamma^2 = \gamma_1 \gamma_1 = \frac{1}{\rho^4} ,$$

$$\gamma = \frac{1}{\rho^2} , \quad (41-52)$$

so that

$$\frac{\partial \gamma}{\partial \rho} = -\frac{2}{\rho^3} = -\frac{2\gamma}{\rho} , \quad (41-53)$$

and

$$C = \frac{\partial \gamma}{\partial \tau} : \frac{\partial \gamma}{\partial \rho} = -\frac{1}{2} \rho \frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} . \quad (41-54)$$

Hence, by (41-50)

$$\rho \frac{\partial}{\partial \rho} = -2 \left(\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} \right)^{-1} \frac{\partial}{\partial \tau} , \quad (41-55)$$

and the boundary condition (41-48) takes the form

$$\frac{\partial T}{\partial \tau} - \frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} T = -\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} f . \quad (41-56)$$

The right-hand side may be transformed as follows. By (41-44) we have

$$f = \Delta W + \underline{m}^T \Delta \underline{g} . \quad (41-57)$$

Let us have a closer look at the vector \underline{m} .

To this effect, let

$$\underline{x} = \underline{x}(\tau) \quad (41-58)$$

be the equation of the isozenithal. Then the vector

$$\underline{e} = \frac{d\underline{x}}{d\tau} \quad (41-59)$$

will be the unit tangent vector of this curve (it will be a unit vector since τ is the arc length). Then

$$\underline{e}^T \text{grad } T = \frac{\partial T}{\partial x_1} \frac{dx_1}{d\tau} = \frac{\partial T}{\partial \tau} \quad (41-60)$$

by the chain rule. Hence there follows from (41-42), (41-47), (41-55) and (41-60):

$$\begin{aligned} \underline{m}^T \text{grad } T &= - \frac{\partial T}{\partial \gamma_1} \gamma_1 = \frac{1}{2} \rho \frac{\partial T}{\partial \rho} \\ &= - \left(\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} \right)^{-1} \frac{\partial T}{\partial \tau} \\ &= - \left(\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} \right)^{-1} \underline{e}^T \text{grad } T . \end{aligned} \quad (41-61)$$

Since the vector $\text{grad } T$ can have any direction, there must be

$$\underline{m}^T = - \left(\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} \right)^{-1} \underline{e}^T . \quad (41-62)$$

Hence the vector \underline{m} is tangent to the isozenithal; since τ is positive upwards, the negative sign implies that \underline{m} is directed downwards.

Thus

$$\underline{m}^T \Delta \underline{g} = - \left(\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} \right)^{-1} \underline{e}^T \Delta \underline{g} . \quad (41-63)$$

Now

$$\underline{e}^T \Delta \underline{g} = - \Delta g' \quad (41-64)$$

is nothing else than the component of the gravity vector $\Delta \underline{g}$ in the downward direction of the isozenithal. Since this direction is very nearly vertical, $\Delta g'$ is almost equal to the usual gravity anomaly Δg in the sense of Molodensky, as entering in eq. (41-67) below.

In view of (41-63) and (41-64), eq. (41-57) becomes

$$f = \Delta W + \left(\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} \right)^{-1} \Delta g' , \quad (41-65)$$

and (41-56) may be written as

$$\frac{\partial T}{\partial \tau} - \frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} T = - \Delta g' - \frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} \Delta W . \quad (41-66)$$

This form of the basic boundary condition is rigorously equivalent to the preceding forms (41-30), (41-43), and (41-48). Though it looks less simple, it is very important because it allows a comparison with the form in which the boundary condition for Molodensky's problem was usually presented earlier. Take, for instance, eq. (8-24b) of (Heiskanen and Moritz, 1967, p.300):

$$\frac{\partial T}{\partial h} - \frac{1}{\gamma} \frac{\partial \gamma}{\partial h} T = -\Delta g \quad (41-67)$$

Here the derivative $\partial/\partial h$ is taken along the normal plumb line. This equation involves certain approximations (cf. *ibid.*, p.85), which are practically permissible but theoretically not rigorous. It was the merit of Krarup to have shown that (41-67) becomes *theoretically exact* if the direction of the normal plumb line is replaced by the direction of the normal isozenithal (the second term on the right-hand side of (41-66) vanishes if the telluroid is defined by $U_Q = W_P$ as usual).

The boundary condition (41-66) is valid on the telluroid Σ , which is a known surface. The problem is to solve Laplace's equation, $\Delta T = 0$, outside Σ with the boundary condition (41-66). Since the isozenithal is, in general, not normal to the surface Σ , we have an *oblique derivative problem*. Such problems are considerably more difficult than boundary-value problems involving normal derivatives, such as Stokes' problem.

42. SPHERICAL APPROXIMATION

If the reference ellipsoid is a nonrotating sphere, then

$$\gamma = \frac{GM}{r^2} \quad (42-1)$$

where G is the gravitational constant, M the total mass, and r the radius vector from the center of the sphere to the point under consideration. The normal gravity vector is then given by

$$\underline{\gamma} = -\gamma \underline{e} \quad (42-2)$$

where

$$\underline{e} = \begin{pmatrix} \cos \phi \cos \lambda \\ \cos \phi \sin \lambda \\ \sin \phi \end{pmatrix} \quad (42-3)$$

denotes the unit vector in the direction of the radius vector, ϕ and λ being geocentric latitude and longitude. The quantities r, ϕ, λ are the usual spherical coordinates.

The cartesian components of $\underline{\gamma}$ may thus be written

$$\begin{aligned} \gamma_1 &= -\frac{GM}{r^2} \cos \phi \cos \lambda, \\ \gamma_2 &= -\frac{GM}{r^2} \cos \phi \sin \lambda, \\ \gamma_3 &= -\frac{GM}{r^2} \sin \phi. \end{aligned} \quad (42-4)$$

The comparison with (41-45) shows that now

$$\rho = r/\sqrt{GM}, \quad (42-5)$$

so that ρ is r apart from a scale factor.

For the non-rotating sphere, the plumb lines, as well as the isozenithals, coincide with the spherical radii. Thus, now

$$\frac{\partial}{\partial \tau} = \frac{\partial}{\partial r}, \quad (42-6)$$

and

$$\frac{1}{\gamma} \frac{\partial \gamma}{\partial \tau} = \frac{1}{\gamma} \frac{\partial \gamma}{\partial r} = -\frac{2}{r} \quad (42-7)$$

by (41-53). Hence (41-66) reduces to

$$\frac{\partial T}{\partial r} + \frac{2}{r} T = -\Delta g' + \frac{2}{r} \Delta W, \quad (42-8)$$

equivalent to (41-48) but with the right-hand side given explicitly.

The boundary-value problem expressed by Laplace's equation

$$\Delta T = 0 \quad (42-9)$$

and the boundary condition (42-8) in spherical coordinates has been called by Krarup the *simple Molodensky problem*; it is the one considered in virtually all practical solutions of the geodetic boundary value problem.

The reason is that, although the reference ellipsoid is not exactly a sphere, its flattening is very small, about 0.3 %, so that on tolerating an error of this order of magnitude in equations relating quantities of the anomalous gravity field, for instance, in the boundary condition, we can formally use spherical boundary condition even in the geodetic case of a reference ellipsoid. This is the so-called *spherical approximation*.

The spherical approximation has been used and described repeatedly; cf. sec. 2 and sec. 39. It may be interpreted geometrically as a mapping of a spatial point P of geodetic coordinates (h, ϕ, λ) , referred to the ellipsoid, into a point P' of spherical coordinates (r, ϕ, λ) , referred to a sphere $r = R$, (Fig.42.1). The radius R of this sphere may be related to the semi-axes a and b of the ellipsoid by

$$R = \sqrt[3]{a^2 b} \quad . \quad (42-10)$$

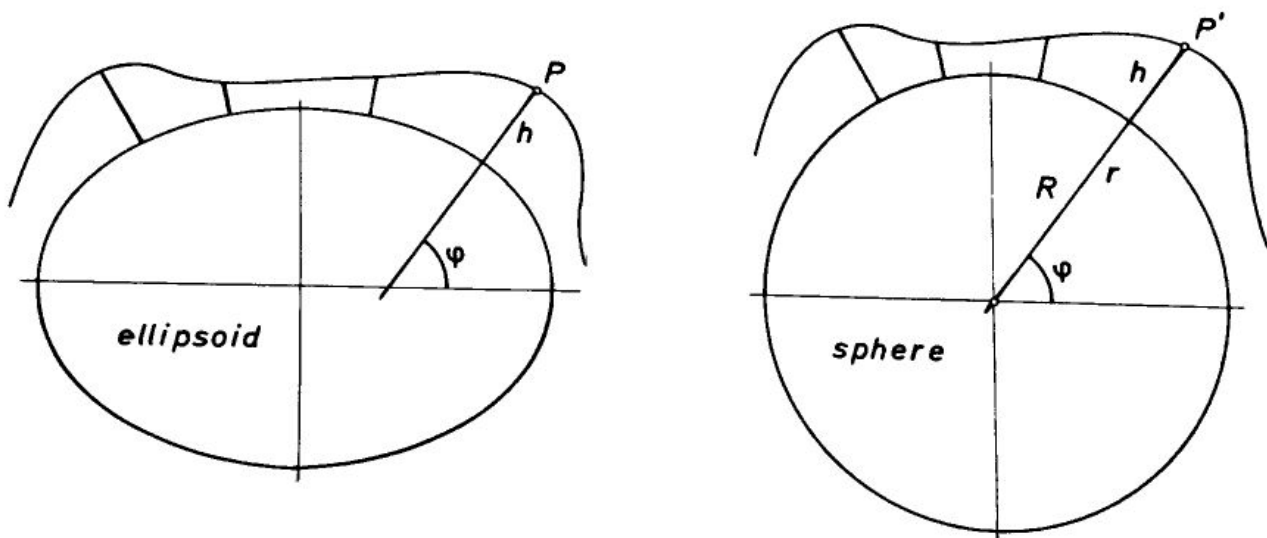


FIGURE 42.1. Spherical approximation as a mapping.

The spherical coordinates ϕ, λ of P' are taken to be equal to the geodetic coordinates of P , and the height h of P' above the sphere is taken equal to the height of P above the ellipsoid; therefore the radius vector r of P' is given by

$$r = R + h. \quad (42-11)$$

The approximation consists in calculating with P' formally as if it were P . As we have repeatedly mentioned, this can be done only with linearized relations involving the anomalous potential T and similar quantities, for which an error on the order of 0.3 % can be neglected. This is usually permissible; therefore, the spherical approximation is used, for instance, in Stokes' formula and in least-squares collocation, and it will also be used in practical solutions of Molodensky's problem, which are based on the "simple Molodensky problem" mentioned above.

The practical boundary condition. In the beginning of sec. 41 we have taken as the telluroid Σ an arbitrary surface approximating the earth's surface S . We shall now specialize Σ in the following way, which is generally used; cf. (Heiskanen and Moritz, 1967, p.292).

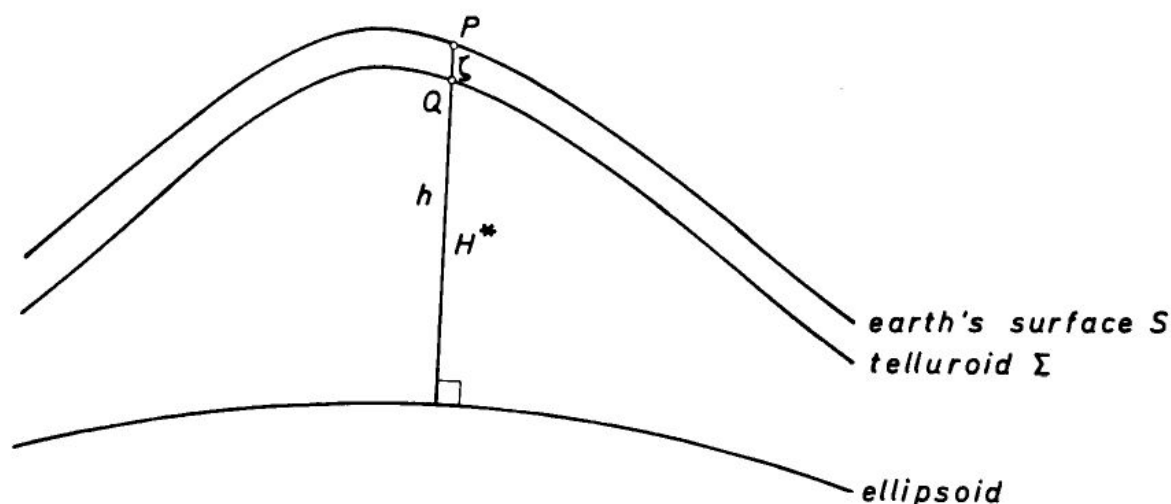


FIGURE 42.2. *The telluroid.*

Consider the ellipsoidal normal through a point P of S (Fig.42.2). On this normal select that (uniquely defined) point Q for which

$$U_Q = W_P ; \quad (42-12)$$

that is, the normal (ellipsoidal) potential U at Q is to be equal to the actual potential W at P . The ellipsoidal height of Q is called the *normal height* H^* of P , and the difference

$$\zeta = h - H^* = QP \quad (42-13)$$

is called the *height anomaly*.

The telluroid defined in this way is similar to the "Marussi telluroid" defined by (41-6) for which also $U_Q = W_P$ but Q and P do not lie on the same ellipsoidal normal. The definition (41-6) is, therefore, geometrically less simple though more rigorous if the astronomical coordinates ϕ, λ are given rather than the geodetic coordinates ϕ, λ of P . Practically, however, the deviation of ϕ_P from $\phi_Q = \phi_P$ according to (41-6), being the deflection of the vertical, is negligible for the present purpose since ϕ and λ of Q are used only for computing normal gravity γ and other quantities of the normal field, which very weakly depend on ϕ and not at all on λ . Therefore, the present definition may very well be used practically, and we shall do so in the sequel.

From $U_Q = W_P$ there follows that the potential anomaly, defined by (41-3), is zero, so that (42-8), on writing $\Delta g' = \Delta g$, reduces to

$$\frac{\partial T}{\partial r} + \frac{2}{r}T = -\Delta g, \quad (42-14)$$

which is a boundary condition given on the telluroid Σ ; the gravity anomaly Δg is defined by

$$\Delta g = g_P - \gamma_Q \quad (42-15)$$

as the difference: gravity on the earth's surface minus normal gravity at the telluroid.

The height anomaly ζ is expressed in terms of T by

$$\zeta = \frac{T}{\gamma}. \quad (42-16)$$

Strictly speaking, γ in this formula refers to Q but we may also take for γ a mean value such as 980 gal.

Eq. (42-16) is a generalization of Bruns' formula (2-31) to the problem of Molodensky and is derived in the same way. In fact, if S denotes the geoid instead of the earth's surface, then the "telluroid" Σ reduces to the ellipsoid, the height anomaly ζ becomes the geoidal height N , and the boundary condition (42-14), with $r = R$ (in the spherical approximation, the ellipsoid is the sphere $r = R$), becomes Stokes' condition (2-33).

The validity of the linearized problem. Let us finally provide a simple reasoning to show that the linearized problem in the sense of sec. 41 is a practically sufficient substitute of the nonlinear Molodensky problem as outlined in sec. 40.

The linearization is performed with respect to T , ζ , Δg and similar quantities of the anomalous gravitational field. The neglected quantities are on the order of

$$\left(\frac{\zeta}{R}\right)^2 \approx \left(\frac{60\text{m}}{6 \cdot 10^6\text{m}}\right)^2 = 10^{-10} \quad (42-17)$$

or

$$\left(\frac{\Delta g}{g}\right)^2 \approx \left(\frac{100\text{mgal}}{10^6\text{mgal}}\right)^2 = 10^{-8} \quad (42-18)$$

(relative error). Such quantities are negligible in view of the present accuracy of gravimetric geodesy, which is usually around 10^{-6} .

43. MOLODENSKY'S SOLUTION

The "simple Molodensky problem" consists in solving Laplace's equation

$$\Delta T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} = 0 \quad (43-1)$$

under the boundary condition (42-14)

$$\frac{\partial T}{\partial r} + \frac{2}{r} T = -\Delta g \quad (43-2)$$

After this boundary-value problem has been solved, the height anomaly ζ is obtained by Bruns' equation (42-16).

It is assumed that the boundary condition (43-2) is satisfied on the given telluroid Σ , and that Laplace's equation (43-1) holds everywhere outside Σ . This assumption is not completely valid if S lies above Σ but it is easily seen that, in keeping with the present linearization, no additional error is introduced in this way.

Molodensky (Molodenskii et al., 1962, sec.V-15) derived a practically useful and elegant solution by representing T as the potential of a surface layer and transforming the boundary condition into an integral equation. We shall derive this integral equation following (Heiskanen and Moritz, 1967, sec.8-6).

We express the anomalous potential T as the potential of a surface layer on the telluroid Σ :

$$T = \iint_{\Sigma} \frac{\phi}{l} d\Sigma = \iint_{\Sigma} \phi l^{-1} d\Sigma \quad (43-3)$$

T refers to a point P , called the "reference point" or "computation point", and l is the distance between P and the surface element $d\Sigma$ (Fig.43.1); ϕ is a function defined on Σ and representing the density of the surface layer.

Since a surface layer potential is harmonic outside Σ , Laplace's equation (43-1) is automatically satisfied. Therefore, we can substitute (43-3) into the boundary condition (43-2). *Outside* Σ we have

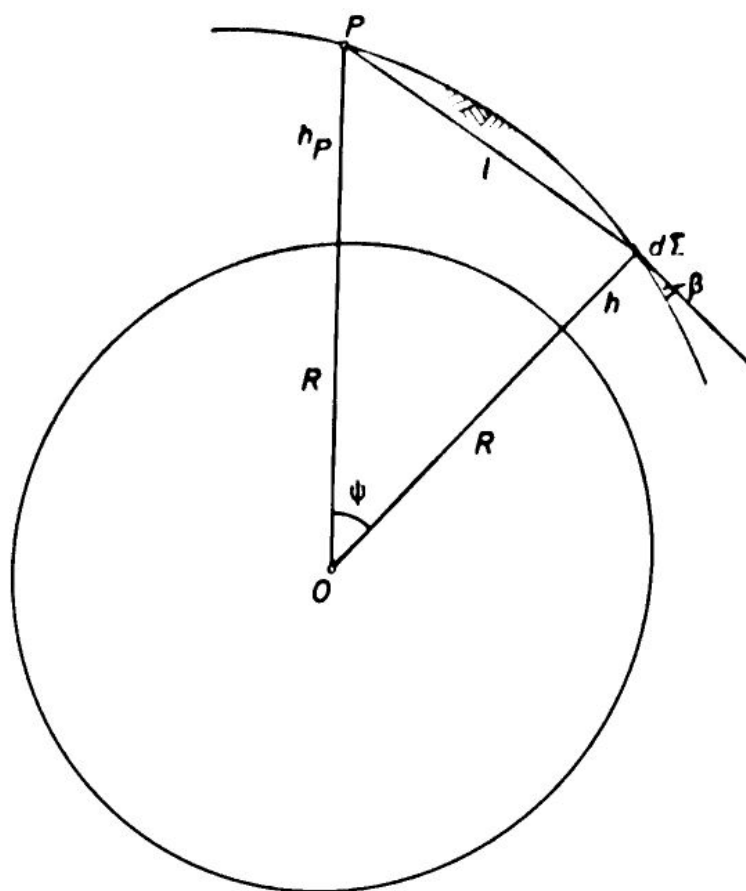
$$\frac{\partial T}{\partial r_P} = \iint_{\Sigma} \phi \frac{\partial l^{-1}}{\partial r_P} d\Sigma, \quad (43-4)$$

but this equation is no longer valid *on* Σ since the derivatives of a surface layer potential are discontinuous at the surface. Instead, we have on S

$$\frac{\partial T}{\partial r_P} = -2\pi\phi_P \cos \beta_P + \iint_{\Sigma} \phi \frac{\partial l^{-1}}{\partial r_P} d\Sigma, \quad (43-5)$$

according to eq. (1-19a) of (Heiskanen and Moritz, 1967, p.6). Thus (43-2) gives

$$2\pi\phi \cos \beta = \iint_{\Sigma} \left(\frac{\partial l^{-1}}{\partial r_P} + \frac{2l^{-1}}{r_P} \right) \phi d\Sigma = \Delta g. \quad (43-6)$$

FIGURE 43.1. *The geometry of Molodensky's solution.*

A basic notational convention. Eq. (43-6) uses a notational convention which will be frequently employed in the sequel. Free terms containing no integral, such as $2\pi\phi\cos\beta$ and Δg , refer to some point; this *reference point* will be the point P in Fig. 43.1. In these free terms, the reference to P is understood without further notational indication. It is more difficult with quantities under the integral sign, which may refer to P or to $d\Sigma$ or to both, such as l . Quantities under the integral sign referring to P will bear the subscript P , for instance r_P ; quantities bearing no subscript will refer to $d\Sigma$. The symbol l will always denote the distance between P and $d\Sigma$. This notation is in agreement with Fig. 43.1: h is the elevation of $d\Sigma$ and h_P is the elevation of P .

This notation for quantities under the integral sign appears natural; the essential notational convention consists in *omitting the subscript P in the "free terms"* as mentioned above. For instance, $\phi\cos\beta$ in (43-6) is the same as $\phi_P\cos\beta_P$ in (43-5). This notational convention appears less natural; it is, however, very useful as the following developments will show.

Transformation of the integral equation. Since the reference ellipsoid is formally considered a sphere, we have by (42-11) and Fig. 43.1.

$$r_p = R + h_p, \quad R = r + h, \quad (43-7)$$

where h is the height above the ellipsoid. As we work with the telluroid, the ellipsoidal height h of a telluroid point is the normal height H^* of the corresponding point of the earth's surface (Fig. 42.2). To the given accuracy we may as well use the orthometric height; in practice, we shall be satisfied to take approximate heights from suitable topographic maps or digital terrain models.

We have

$$l = \sqrt{r_p^2 + r^2 - 2r_p r \cos \psi}, \quad (43-8)$$

which on differentiation gives

$$\frac{\partial l^{-1}}{\partial r_p} = - \frac{r_p - r \cos \psi}{l^3}. \quad (43-9)$$

A simple calculation shows that

$$\frac{\partial l^{-1}}{\partial r_p} + \frac{2l^{-1}}{r_p} = \frac{3}{2r_p l} + \frac{r^2 - r_p^2}{2r_p l^3}. \quad (43-10)$$

Thus (43-6) takes the form

$$2\pi \phi \cos \beta - \iint_{\Sigma} \left(\frac{3}{2r_p l} + \frac{r^2 - r_p^2}{2r_p l^3} \right) \phi d\Sigma = \Delta g. \quad (43-11)$$

The surface element $d\Sigma$ may be eliminated by noting that the projection of $d\Sigma$ onto the local horizon is given by

$$d\Sigma \cos \beta.$$

This is also equal to

$$r^2 d\sigma,$$

where $d\Omega$ is the element of solid angle, because r is the radius vector of $d\Sigma$. Hence we have

$$d\Sigma = r^2 \sec \beta d\sigma. \quad (43-12)$$

Thus the integral in (43-11) can be extended over the unit sphere σ :

$$2\pi \phi \cos \beta - \iiint_{\sigma} \left(\frac{3}{2l} + \frac{r^2 - r_P^2}{2l^3} \right) \frac{r^2}{r_P} \sec \beta \cdot \phi d\sigma = \Delta g. \quad (43-13)$$

This is the basic integral equation for the simple Molodensky problem. *Planar approximation.* We note that

$$r = R + h = R \left(1 + \frac{h}{R} \right) \quad (43-14)$$

differs from R by less than 10^{-3} , which is smaller than the error of the spherical approximation. Thus we may safely put

$$\begin{aligned} \frac{r^2}{r_P} &= R, \\ r^2 - r_P^2 &= (h - h_P)(r + r_P) = 2R(h - h_P), \end{aligned}$$

obtaining

$$2\pi \phi \cos \beta - \iiint_{\sigma} \left(\frac{3R}{2l} + \frac{R^2(h - h_P)}{l^3} \right) \sec \beta \cdot \phi d\sigma = \Delta g. \quad (43-15)$$

This equation is simpler than (43-13), but hardly less accurate.

We can also simplify the expression for the distance l . We find

$$\begin{aligned} l^2 &= r_P^2 + r^2 - 2r_P r \cos \psi \\ &= (R + h_P)^2 + (R + h)^2 - 2(R + h_P)(R + h) \cos \psi \\ &= 2R^2(1 - \cos \psi) + 2R(h + h_P)(1 - \cos \psi) + h_P^2 + h^2 - 2h_P h \cos \psi \\ &= 4R^2 \sin^2 \frac{\psi}{2} \left(1 + \frac{h + h_P}{R} + \frac{h_P h}{R^2} \right) + (h - h_P)^2. \end{aligned}$$

For the same reason as above we may neglect $(h + h_p)/R$ and $h_p h/R^2$, obtaining

$$l^2 = l_0^2 + (h - h_p)^2, \quad (43-16)$$

$$l = l_0 \sqrt{1 + \left(\frac{h - h_p}{l_0} \right)^2}, \quad (43-17)$$

where

$$l_0 = 2R \sin \frac{\psi}{2}. \quad (43-18)$$

This procedure, neglecting a relative error of

$$\frac{h}{R} < \frac{8 \text{ km}}{6371 \text{ km}} \doteq 0.001 = 0.1\%,$$

is called the planar approximation. The name recalls that disregarding a term h/R is equivalent to letting $R \rightarrow \infty$ in this term, that is, to a formal transition from the sphere to a plane. In fact, also (43-16) can be interpreted in this way (Fig. 43.2).

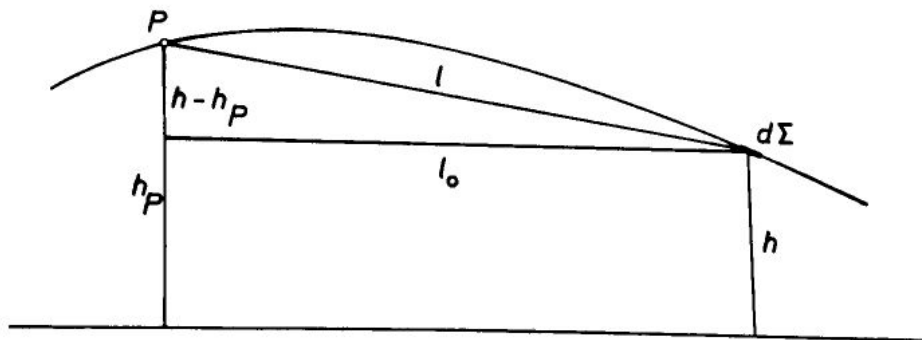


FIGURE 43.2. The planar approximation.

Note, however, that this provides only an illustrative interpretation of a formal procedure; it does not mean that a plane is now used as a geometric reference surface. The "real" reference surface always remains the ellipsoid.

The planar approximation is practically justified, as well as the spherical approximation; both are employed in the standard solutions of Molodensky's problem.

The Molodensky shrinking. The solution of the basic integral equation (43-15) will be found by a series expansion with respect to a certain parameter k . It is convenient first to introduce, instead of the function ϕ , a new auxiliary function

$$\chi = \phi \sec \beta$$

(43-19)

with the result

$$2\pi\chi(1+\tan^2\beta)^{-1} - \iint \left[\frac{3R}{2l} + \frac{R^2(h-h_P)}{l^3} \right] \chi d\sigma = \Delta g, \quad (43-20)$$

omitting σ below the double integral sign as self-evident. In order to solve this equation, we apply the following artifice ("Molodensky shrinking"¹): we replace h by kh and $\tan\beta$ by $k\tan\beta$, where k is a parameter with $0 \leq k \leq 1$. The integral equation (43-20) then becomes

$$2\pi\chi(1+k^2\tan^2\beta)^{-1} - \frac{3}{2}R\iint \frac{\chi}{l_k} d\sigma - R^2\iint \frac{k(h-h_P)}{l_k^3} \chi d\sigma = \Delta g, \quad (43-21)$$

where

$$l_k^2 = l_0^2 + k^2(h-h_P)^2 = l_0^2 \left[1 + k^2 \left(\frac{h-h_P}{l_0} \right)^2 \right], \quad (43-22)$$

l_0 being given by (43-18).

Now we can expand in series of powers of k :

$$\frac{1}{l_k} = \frac{1}{l_0} \left[1 + \sum_{r=1}^{\infty} a_r k^{2r} \left(\frac{h-h_P}{l_0} \right)^{2r} \right], \quad (43-23)$$

¹The geometrical interpretation is a shrinking of the topography by the factor k . For instance, if $k = 0.1$, all elevations are reduced to 1/10 of their original size.

$$\frac{1}{l_k^3} = \frac{1}{l_0^3} \left[1 + \sum_{r=1}^{\infty} b_r k^{2r} \left(\frac{h-h_p}{l_0} \right)^{2r} \right] \quad (43-24)$$

with

$$a_r = \begin{pmatrix} -1/2 \\ r \end{pmatrix}, \quad b_r = \begin{pmatrix} -3/2 \\ r \end{pmatrix} \quad (43-25)$$

being binomial coefficients, and

$$(1 + k^2 \tan^2 \beta)^{-1} = 1 + \sum_{r=1}^{\infty} (-1)^r k^{2r} \tan^{2r} \beta. \quad (43-26)$$

Finally we expand also the unknown function χ :

$$\chi = \sum_{n=0}^{\infty} k^n \chi_n. \quad (43-27)$$

These series are all substituted in (43-21):

$$\begin{aligned} & 2\pi \left(\chi_0 + k\chi_1 + k^2\chi_2 + \dots \right) \left(1 - k^2 \tan^2 \beta + k^4 \tan^4 \beta \dots \right) - \\ & - \frac{3}{2} R \iint \frac{1}{l_0} \left(\chi_0 + k\chi_1 + k^2\chi_2 + \dots \right) \left(1 - \frac{1}{2} k^2 n^2 + \frac{3}{8} k^4 n^4 \dots \right) d\sigma - \\ & - R^2 \iint \frac{k n}{l_0^2} \left(\chi_0 + k\chi_1 + k^2\chi_2 \dots \right) \left(1 - \frac{3}{2} k^2 n^2 + \frac{15}{8} k^4 n^4 \dots \right) d\sigma - \\ & - \Delta g = 0, \end{aligned} \quad (43-28)$$

where we have put

$$n = \frac{h-h_p}{l_0}. \quad (43-29)$$

If we carry out the multiplications and combine those terms that are multiplied by the same power of k we get

$$2\pi\chi_0 - \frac{3}{2} R \iint \frac{\chi_0}{l_0} d\sigma - \Delta g +$$

$$\begin{aligned}
& + k \left(2\pi x_1 - \frac{3}{2} R \iint \frac{x_1}{l_0} d\sigma - R^2 \iint \frac{\eta}{l_0^2} x_0 d\sigma \right) + \\
& + k^2 \left(2\pi x_2 - \frac{3}{2} R \iint \frac{x_2}{l_0} d\sigma - R^2 \iint \frac{\eta}{l_0^2} x_1 d\sigma + \frac{3R}{4} \iint \frac{\eta^2}{l_0} x_0 d\sigma - 2\pi x_0 \tan^2 \beta \right) + \\
& + \dots = 0 .
\end{aligned} \tag{43-30}$$

This equation is identically satisfied if all coefficients of k^n for $n = 0, 1, 2, 3, \dots$ are set equal to zero. This gives the following system of integral equations:

$$2\pi x_n - \frac{3}{2} R \iint \frac{x_n}{l_0} d\sigma = G_n \tag{43-31}$$

where

$$\begin{aligned}
G_0 &= \Delta g , \\
G_1 &= R^2 \iint \frac{h-h_P}{l_0^3} x_0 d\sigma , \\
G_2 &= R^2 \iint \frac{h-h_P}{l_0^3} x_1 d\sigma - \frac{3R}{4} \iint \frac{(h-h_P)^2}{l_0^3} x_0 d\sigma + 2\pi x_0 \tan^2 \beta , \\
G_3 &= R^2 \iint \frac{h-h_P}{l_0^3} x_2 d\sigma - \frac{3R}{4} \iint \frac{(h-h_P)^2}{l_0^3} x_1 d\sigma - \\
& - \frac{3}{2} R^2 \iint \frac{(h-h_P)^3}{l_0^5} x_0 d\sigma + 2\pi x_1 \tan^2 \beta , \\
& \dots
\end{aligned} \tag{43-32}$$

We are constantly using the abbreviation

$$\iint = \iint_{\sigma} . \tag{43-33}$$

The case $n = 0$. For $n = 0$, eq. (43-31) becomes

$$2\pi x_0 - \frac{3}{2} R \iint \frac{x_0}{l_0} d\sigma = \Delta g . \tag{43-34}$$

This case corresponds to $h = 0$, that is, to the telluroid coinciding with the reference "sphere". For this case, (43-3) reduces to

$$T_0 = \iint \frac{\phi_0}{l_0} R^2 d\sigma = R^2 \iint \frac{x_0}{l_0} d\sigma \quad (43-35)$$

since now $\beta = 0$ in (43-19). The substitution of (43-35) into the preceding equation gives

$$x_0 = \frac{1}{2\pi} \left(\Delta g + \frac{3}{2R} T_0 \right), \quad (43-36)$$

and on expressing T_0 by Stokes' formula (2-35),

$$x_0 = \frac{1}{2\pi} \Delta g + \frac{3}{16\pi^2} \iint \Delta g S(\psi) d\sigma, \quad R^2 \iint \frac{x_0}{l_0} d\sigma = \frac{R}{4\pi} \iint G_0 S(\psi) d\sigma. \quad (43-37)$$

The series solution. This formula may be used to solve (43-31) since (43-34), on replacing x_0 by x_n and Δg by G_n , becomes (43-31). Hence the solution of (43-31) is simply

$$x_n = \frac{1}{2\pi} G_n + \frac{3}{16\pi^2} \iint G_n S(\psi) d\sigma, \quad R^2 \iint \frac{x_n}{l_0} d\sigma = \frac{R}{4\pi} \iint G_n S(\psi) d\sigma, \quad (43-38)$$

where $S(\psi)$ is Stokes' function.

Thus we get x_0 from (43-38) with $G_0 = \Delta g$, then we find G_1 from (43-32), after that we get x_1 from (43-38), then G_2 from (43-32) and x_2 from (43-38), and so forth.

Finally we determine the anomalous potential by (43-3). Using (43-12), (43-19), (43-23), (43-27), and (43-29) we get as a planar approximation:

$$\begin{aligned} T &= \iint_{\sigma} x \cos \beta \cdot l^{-1} \cdot r^2 \sec \beta d\sigma = R^2 \iint x l^{-1} d\sigma \\ &= R^2 \iint \frac{1}{l_0} \left(x_0 + kx_1 + k^2 x_2 + \dots \right) \left(1 - \frac{1}{2} k^2 n^2 + \frac{3}{8} k^4 n^4 \dots \right) d\sigma \\ &= T_0 + kT_1 + k^2 T_2 + \dots \end{aligned} \quad (43-39)$$

In view of (43-38) we obtain

$$T_0 = \frac{R}{4\pi} \iint G_0 S(\psi) d\sigma,$$

$$T_1 = \frac{R}{4\pi} \iint G_1 S(\psi) d\sigma, \quad (43-40)$$

$$T_2 = \frac{R}{4\pi} \iint G_2 S(\psi) d\sigma - \frac{R^2}{2} \iint \frac{(h-h_p)^2}{l_0^3} x_0 d\sigma,$$

$$T_3 = \frac{R}{4\pi} \iint G_3 S(\psi) d\sigma - \frac{R^2}{2} \iint \frac{(h-h_p)^2}{l_0^3} x_1 d\sigma,$$

. . . .

The parameter k has served only as a tool to get a convenient mechanism for a series expansion; at the end, of course, we set $k = 1$ (since this corresponds to the actual earth's surface) to get

$$T = T_0 + T_1 + T_2 + \dots = \sum_{n=0}^{\infty} T_n. \quad (43-41)$$

Eq. (43-41), together with (43-32), (43-38), and (43-40), constitutes the solution known as *Molodensky's series*; it was derived by M.S. Molodensky in 1960 (Molodenskii et al., 1962).

The height anomaly ζ is then obtained by Bruns' equation (42-16).

To first order, this solution coincides with the solution described in sec. 8-7 of (Heiskanen and Moritz, 1967).

We finally mention that the relation between Δg and T_0 (that is, between Δg and T in Stokes' problem) is given by the simple Stokes' formula (2-35) only if the anomalous gravitational field does not contain a spherical-harmonic term of degree zero, which means that the mass of the reference ellipsoid equals the mass of the earth. If this is not the case, then Stokes' formula must be slightly generalized (Heiskanen and Moritz, 1967, sec.2-19):

$$T_0 = \frac{R}{4\pi} \iint \Delta g [S(\psi) - 1] d\sigma. \quad (43-42)$$

Thus it would be more correct to replace in (43-37), (43-38), and (43-40) the function $S(\psi)$ by $S(\psi) - 1$. However, the mass of the earth is now very well known; hence we may presuppose that the mass of the reference ellipsoid is taken equal to the earth's mass, and use $S(\psi)$ as we did in this section. A related discussion will be found in the following section.

44. BROVAR'S SOLUTION

Molodensky represented the anomalous potential T as the potential of a surface layer (43-3):

$$T = \iint_{\Sigma} \phi l^{-1} d\Sigma \quad (44-1)$$

The reason why such a representation is possible is that l^{-1} is harmonic as a function of the reference point P , satisfying Laplace's equation

$$\Delta(l^{-1}) = 0 \quad (44-2)$$

outside Σ . It follows that

$$\Delta T = \iint_{\Sigma} \phi \Delta(l^{-1}) d\Sigma = 0, \quad (44-3)$$

so that T is, in fact, harmonic outside Σ .¹

Brovar's (1964) idea is to replace l^{-1} by a different harmonic function E , arriving at

$$T = \iint_{\Sigma} \lambda E d\Sigma \quad (44-4)$$

This representation is valid since T will be harmonic if E is, by the same argument as used for l^{-1} . We may regard (44-4) as the potential of a generalized surface layer, and regard the function λ defined on Σ as a generalized surface density.

Then, at a point P outside the surface Σ , we have by (43-2):

$$\Delta g_P = -\frac{\partial T}{\partial r_P} - \frac{2}{r_P} T = \iint_{\Sigma} \lambda \left(-\frac{\partial E}{\partial r_P} - \frac{2}{r_P} E \right) d\Sigma \quad (44-5)$$

¹We have interchanged the order of the operator Δ and the integral, which would have to be justified for full mathematical rigor. In these sections, we are frequently proceeding in this way; the mathematically minded reader is invited to supply the justifications himself. In the present case, the justification is easy: if the point P , to which T refers, lies outside Σ , then l^{-1} is a regular function, and the interchangeability follows from differentiating an ordinary definite integral with respect to a parameter.

The function E may be selected in such a way that the kernel

$$K = -\frac{\partial E}{\partial r_P} - \frac{2}{r_P} E \quad (44-6)$$

has a suitable form.

Of particular advantage is the choice

$$E = \frac{1}{4\pi} \sum_{n=0}^{\infty} \frac{2n+1}{n-1} \frac{r^n}{r_P^{n+1}} P_n(\cos \psi) . \quad (44-7)$$

Thus E is a function of two points P and Q , where P is the reference point as usual and Q is the point at which the surface element $d\Sigma$ is situated; r denotes the radius vector of Q and ψ is the angle between r_P and r . The prime (') after the summation sign means that the sum does not contain a term with $n = 1$. The spherical-harmonic representation (44-7) shows directly that E is harmonic as a function of P .

The series in (44-7) possesses a sum which is related to Stokes function; in fact, for $r = r_P$ we essentially have Stokes' function; cf. (Heiskanen and Moritz, 1967, eq.(2-169)). The summation may be effected by the methods of sec. 23. Putting

$$\frac{r}{r_P} = \sigma \quad (44-8)$$

and noting that

$$\frac{2n+1}{n-1} = 2 + \frac{3}{n-1} , \quad (44-9)$$

we see that (44-7) is a linear combination of the functions $F(\sigma, \psi)$ and $F_{-1}(\sigma, \psi)$ as given by eqs. (23-36), (23-42), and (23-50).

The result may be formulated as follows. The *generalized Stokes function*, defined as

$$S(r_P, \psi, r) = \sum_{n=2}^{\infty} \frac{2n+1}{n-1} \frac{r^n}{r_P^{n+1}} P_n(\cos \psi) , \quad (44-10)$$

has the closed expression (*ibid.*, eq.(2-162))

$$\begin{aligned}
 S(r_P, \psi, r) = & \frac{2}{1} + \frac{1}{r_P} - \frac{31}{r_P^2} - \frac{5r}{r_P^2} \cos \psi - \\
 & - \frac{3r}{r_P^2} \cos \psi \ln \frac{r_P - r \cos \psi + 1}{2r_P} .
 \end{aligned} \quad (44-11)$$

Thus (44-7) takes the form

$$E = \frac{1}{4\pi} \left[S(r_P, \psi, r) - \frac{1}{r_P} \right] \quad (44-12)$$

(the term $1/r_P$ comes from $n = 0$), and (44-4) becomes

$$T = \frac{1}{4\pi} \iint_{\Sigma} \lambda \left[S(r_P, \psi, r) - \frac{1}{r_P} \right] d\Sigma . \quad (44-13)$$

To compute the kernel K , we substitute (44-12) with (44-11) into (44-6) and perform the differentiation with respect to r_P . The result is simply

$$K(r_P, \psi, r) = -\frac{1}{4\pi} \frac{r^2 - r_P^2}{r_P l^3} - \frac{3}{4\pi} \frac{r}{r_P^3} \cos \psi , \quad (44-14)$$

so that, on omitting the small last term, (44-5) becomes

$$\frac{\partial T}{\partial r_P} + \frac{2}{r_P} T = \frac{1}{4\pi} \iint_{\Sigma} \frac{r^2 - r_P^2}{r_P l^3} \lambda d\Sigma . \quad (44-15)$$

This formula is valid as long as the point P lies outside the telluroid Σ . It no longer holds at Σ because the main singularity of the function $S(r_P, \psi, r)$ is that of $2l^{-1}$ as $l \rightarrow 0$, according to (44-11). Hence, on transition from the outside of Σ to Σ itself, the function E behaves as

$$\frac{1}{2\pi} l^{-1} ,$$

so that the integral (44-4) behaves as

$$\frac{1}{2\pi} \iint \lambda l^{-1} d\Sigma , \quad (44-16)$$

that is, as a surface layer potential: the radial derivative $\partial/\partial r_P$ of (44-13) will undergo a jump

$$- \lambda \cos \beta . \quad (44-17)$$

This follows from (43-4) and (43-5) with λ taking the place of $2\pi\phi$, cf. (44-16) and (43-3). Hence (44-15) becomes on Σ :

$$\frac{\partial T}{\partial r_P} + \frac{2}{r_P} T = - \lambda \cos \beta + \frac{1}{4\pi} \iint_{\Sigma} \frac{r^2 - r_P^2}{r_P^3} \lambda d\Sigma . \quad (44-18)$$

It may be shown that the second, logarithmic, singularity of (44-11) does not give rise to a discontinuity of $\partial T/\partial r_P$ on Σ ; cf. (Moritz, 1968a, p.47).

Brovar's integral equation. As the left-hand side of (44-18) is equal to $-\Delta g$, we obtain

$$\lambda \cos \beta - \frac{1}{4\pi} \iint_{\Sigma} \frac{r^2 - r_P^2}{r_P^3} \lambda d\Sigma = \Delta g . \quad (44-19)$$

This is the desired integral equation for determining the density λ from the given gravity anomaly Δg on Σ .

In this equation, the subscript P has been used only within the integral, in agreement with our basic notational convention (sec.43). Note, e.g., that λ in the first term in (44-19) refers to P , whereas λ under the integral refers to the point Q at which $d\Sigma$ is situated.

Brovar's integral equation (44-19) is essentially simpler than Molodensky's integral equation (43-11) since for the telluroid coinciding with a sphere (for $r = r_P = R$) the integral in (44-19) vanishes, which is not the case with (43-11).

The further treatment is quite analogous to Molodensky's method outlined in sec. 43. By means of (43-12) we replace the integration over the telluroid by an integration over the unit sphere, getting

$$\lambda \cos \beta - \frac{1}{4\pi} \iint_{\sigma} \frac{r^2 - r_P^2}{r_P^3} r^2 \sec \beta \lambda d\sigma = \Delta g . \quad (44-20)$$

Then we introduce as a new auxiliary function

$$\mu = \lambda \sec \beta \quad (44-21)$$

to obtain

$$\mu \cos^2 \beta - \frac{1}{4\pi} \iint_{\sigma} \frac{r^2 - r_P^2}{r_P l^3} r^2 \mu d\sigma = \Delta g. \quad (44-22)$$

This equation is rigorous for the simple Molodensky problem, except for the missing last term in (44-14) which could, however, easily be added.

Planar approximation and iterative solution. As a planar approximation (sec.43) this last term is indeed zero and we have

$$\frac{r^2 - r_P^2}{r_P l^3} r^2 = \frac{h - h_P}{l^3} \frac{r + r_P}{r_P} r^2 = 2R^2 \frac{h - h_P}{l^3},$$

so that (44-22) reduces to

$$\mu \cos^2 \beta - \frac{R^2}{2\pi} \iint_{\sigma} \frac{h - h_P}{l^3} \mu d\sigma = \Delta g. \quad (44-23)$$

Since the integral is zero for the sphere, it can be expected to be relatively small for the telluroid which (on a global scale) differs little from a sphere. Thus the integral equation (44-23) lends itself to an iterative solution:

$$\begin{aligned} \mu^{(1)} &= \Delta g \sec^2 \beta, \\ \mu^{(2)} &= \sec^2 \beta \left[\Delta g + \frac{R^2}{2\pi} \iint_{\sigma} \frac{h - h_P}{l^3} \mu^{(1)} d\sigma \right], \\ &\dots \\ \mu^{(i+1)} &= \sec^2 \beta \left[\Delta g + \frac{R^2}{2\pi} \iint_{\sigma} \frac{h - h_P}{l^3} \mu^{(i)} d\sigma \right], \\ &\dots \end{aligned} \quad (44-24)$$

This method will be used in sec. 47 in a discussion of convergence.

Use of Molodensky's shrinking. Practically more important seems to be an expansion with respect to Molodensky's parameter k . We shall proceed in full analogy to sec. 43.

On replacing h by kh and $\tan \beta$ by $k \tan \beta$, the integral equation (44-23) becomes

$$\mu(1+k^2 \tan^2 \beta)^{-1} - \frac{R^2}{2\pi} \iint_{\sigma} \frac{k(h-h_p)}{l_k^3} \mu d\sigma = \Delta g; \quad (44-25)$$

there is now, by (43-22),

$$l_k^2 = l_0^2 + k^2(h-h_p)^2 = l_0^2(1+k^2 n^2), \quad (44-26)$$

where again

$$n = \frac{h-h_p}{l_0} \quad \text{and} \quad l_0 = 2R \sin \frac{\psi}{2}. \quad (44-27)$$

We expand the quantities depending on k into power series (43-24) and (43-26):

$$\frac{1}{l_k^3} = \frac{1}{l_0^3} \sum_{r=0}^{\infty} b_r k^{2r} n^{2r}, \quad (44-28)$$

$$(1+k^2 \tan^2 \beta)^{-1} = \sum_{r=0}^{\infty} (-1)^r k^{2r} \tan^{2r} \beta. \quad (44-29)$$

Finally we also expand the unknown function μ :

$$\mu = \sum_{p=0}^{\infty} k^p \mu_p. \quad (44-30)$$

These series are all substituted into (44-25):

$$\begin{aligned} & \sum_{p=0}^{\infty} k^p \mu_p \sum_{r=0}^{\infty} (-1)^r k^{2r} \tan^{2r} \beta - \\ & - \frac{R^2}{2\pi} \iint_{\sigma} \frac{k n}{l_0^2} \sum_{p=0}^{\infty} k^p \mu_p \sum_{r=0}^{\infty} b_r k^{2r} n^{2r} d\sigma = \Delta g. \end{aligned} \quad (44-31)$$

A slight transformation gives

$$\begin{aligned} & \sum_{p=0}^{\infty} \sum_{r=0}^{\infty} k^{p+2r} (-1)^r \mu_p \tan^{2r} \beta - \\ & - \sum_{p=0}^{\infty} \sum_{r=0}^{\infty} k^{p+2r+1} b_r \cdot \frac{R^2}{2\pi} \iint_{\sigma} \mu_p \frac{\eta^{2r+1}}{l_0^2} d\sigma = \Delta g . \end{aligned} \quad (44-32)$$

From the theory of multiplication of power series (cf. Knopp, 1964, p.181) it is well known that for series of powers of the variable z there holds

$$\sum_{p=0}^{\infty} a_p z^p \sum_{q=0}^{\infty} b_q z^q = \sum_{n=0}^{\infty} c_n z^n ,$$

where

$$c_n = a_n b_0 + a_{n-1} b_1 + \dots + a_0 b_n = \sum_{s=0}^n a_{n-s} b_s .$$

This we may write in the form

$$\sum_{p=0}^{\infty} \sum_{q=0}^{\infty} a_p b_q z^{p+q} = \sum_{n=0}^{\infty} \sum_{s=0}^n a_{n-s} b_s z^n . \quad (44-33)$$

We shall apply this formula to the first sum in (44-32), putting

$$q = 2r , \quad s = 2r , \quad z = k , \quad (44-34)$$

(terms with odd q and s are zero) and then to the second sum, putting

$$q = 2r+1 , \quad s = 2r+1 , \quad z = k . \quad (44-35)$$

(terms with even q and s are zero). The result is

$$\begin{aligned} & \sum_{n=0}^{\infty} k^n \sum_{r=0}^M (-1)^r \mu_{n-2r} \tan^{2r} \beta - \\ & - \sum_{n=0}^{\infty} k^n \sum_{r=0}^N b_r \cdot \frac{R^2}{2\pi} \iint_{\sigma} \mu_{n-2r-1} \frac{\eta^{2r+1}}{l_0^2} d\sigma - \Delta g = 0 . \end{aligned} \quad (44-36)$$

The upper limits M and N of summation are taken in such a way that the subscript of μ never becomes negative, hence

$$M = \begin{cases} \frac{n}{2} & \text{if } n \text{ is even,} \\ \frac{n-1}{2} & \text{if } n \text{ is odd,} \end{cases} \quad (44-37)$$

$$N = \begin{cases} \frac{n-2}{2} & \text{if } n \text{ is even,} \\ \frac{n-1}{2} & \text{if } n \text{ is odd.} \end{cases}$$

It is clear that r , as an integer summation variable, has nothing to do with the radius vector which is also denoted by r .

Eq. (44-36) is identically satisfied if the sum of all terms multiplied by the same power k^n is zero. This gives for $n = 0$:

$$\mu_0 = \Delta g, \quad (44-38)$$

and for $n > 0$:

$$\sum_{r=0}^M (-1)^r \mu_{n-2r} \tan^{2r} \beta - \sum_{r=0}^N b_r \cdot \frac{R^2}{2\pi} \iint_{\sigma} \mu_{n-2r-1} \frac{n}{l_0^2} d\sigma = 0. \quad (44-39)$$

This equation can be solved for μ_n (this term is obtained by putting $r = 0$ in the first summand):

$$\mu_n = \sum_{r=0}^N b_r \frac{R^2}{2\pi} \iint_{\sigma} \frac{(h-h_P)^{2r+1}}{l_0^{2r+3}} \mu_{n-2r-1} d\sigma - \sum_{r=1}^M (-1)^r \mu_{n-2r} \tan^{2r} \beta. \quad (44-40)$$

This equation expresses μ_n in terms of $\mu_0, \mu_1, \dots, \mu_{n-1}$ and thus allows the consecutive computation of μ_n , starting with $\mu_0 = \Delta g$ by (44-38).

The anomalous potential. We finally find T by (44-13), which may also be expanded as a power series with respect to k .

On substituting (43-12) and (44-21) we have

$$T = \frac{1}{4\pi} \iint_{\sigma} \nu \left[S(r_p, \psi, r) - \frac{1}{r_p} \right] R^2 d\sigma. \quad (44-41)$$

For $h = 0$, the telluroid coinciding with the sphere $r = R$, this becomes

$$T_0 = \frac{R^2}{4\pi} \iint_{\sigma} \Delta g \left[S(R, \psi, R) - \frac{1}{R} \right] d\sigma, \quad (44-42)$$

since $\nu_0 = \Delta g$ by (44-38). The definition (44-10) of the generalized Stokes' function shows that

$$S(R, \psi, R) = \frac{1}{R} \sum_{n=2}^{\infty} \frac{2n+1}{n-1} P_n(\cos \psi) = R^{-1} S(\psi), \quad (44-43)$$

in view of the well-known expansion of $S(\psi)$ as a series of Legendre functions (cf. Heiskanen and Moritz, 1967, p.97); $S(\psi)$ is the ordinary function of Stokes as given by eq. (2-38). Thus eq. (44-42), for the "zeroth approximation", reduces to

$$T_0 = \frac{R}{4\pi} \iint_{\sigma} \Delta g [S(\psi) - 1] d\sigma. \quad (44-44)$$

This is Stokes' formula extended to the case that the mass M' enclosed by the reference ellipsoid is not equal to the mass M of the earth; cf. eq. (2-189') of (Heiskanen and Moritz, 1967, p.103). If $M' = M$, as is usually supposed, then (44-44) and the usual Stokes' formula (2-35) give the same result since

$$\iint_{\sigma} \Delta g d\sigma = 0 \quad (44-45)$$

in this case.¹

¹This is rigorously true for Stokes' problem, in which Δg refers to the sphere, but can be expected to hold approximately also for Molodensky's problem in which Δg refers to the telluroid.

Let us now consider the difference

$$\Delta S = S(r_p, \psi, r) - S(R, \psi, R) . \quad (44-46)$$

The substitution of (44-11) gives

$$\Delta S = \frac{2}{1} - \frac{2}{1_0} + O(r_p^{-1}) - O(R^{-1}) , \quad (44-47)$$

where $O(r_p^{-1})$ denotes terms that go to zero for $r_p \rightarrow \infty$ as r_p^{-1} . The planar approximation is equivalent to a formal transition $R \rightarrow \infty$, $r_p \rightarrow \infty$, so that (44-47) reduces to

$$\Delta S = \frac{2}{1} - \frac{2}{1_0} \quad (44-48)$$

as a planar approximation. Then (44-46) gives, to the same approximation,

$$S(r_p, \psi, r) = S(R, \psi, R) + \Delta S = \frac{1}{R} S(\psi) + \frac{2}{1} - \frac{2}{1_0} \quad (44-49)$$

and, since $r_p^{-1} \doteq R^{-1}$,

$$S(r_p, \psi, r) - \frac{1}{r_p} = \frac{1}{R} [S(\psi) - 1] + \frac{2}{1} - \frac{2}{1_0} . \quad (44-50)$$

By (43-23) we have, on introducing k so that $1 = 1_k$,

$$\frac{2}{1} - \frac{2}{1_0} = \frac{2}{1_0} \sum_{r=1}^{\infty} a_r k^{2r} n^{2r} . \quad (44-51)$$

On substituting (44-30), (44-50), and (44-51), eq. (44-41) becomes

$$\begin{aligned} T &= \frac{R}{4\pi} \iint_{\sigma} \sum_{p=0}^{\infty} k^p \mu_p \left[S(\psi) - 1 + \frac{2R}{1_0} \sum_{r=1}^{\infty} a_r k^{2r} n^{2r} \right] d\sigma \\ &= \sum_{n=0}^{\infty} k^n \frac{R}{4\pi} \iint_{\sigma} \mu_n [S(\psi) - 1] d\sigma + \\ &\quad + \sum_{p=0}^{\infty} \sum_{r=1}^{\infty} k^{p+2r} a_r \cdot \frac{R^2}{2\pi} \iint_{\sigma} \mu_p \frac{(h-h_p)^{2r}}{1_0^{2r+1}} d\sigma . \end{aligned} \quad (44-52)$$

To the last sum we apply again (44-33) with (44-34). The result is

$$T = \sum_{n=0}^{\infty} k^n T_n \quad (44-53)$$

where

$$T_0 = \frac{R}{4\pi} \iint_{\sigma} \mu_0 [S(\psi) - 1] d\sigma, \quad (44-54)$$

$$T_n = \frac{R}{4\pi} \iint_{\sigma} \mu_n [S(\psi) - 1] d\sigma + \sum_{r=1}^M a_r \frac{R^2}{2\pi} \iint_{\sigma} \frac{(h-h_p)^{2r}}{r^{2r+1}} \mu_{n-2r} d\sigma \quad (44-55)$$

for $n > 0$; the integer M is defined by (44-37).

The coefficients a_r in (44-55) and b_r in (44-40) are given by (43-25); more explicitly we have

$$a_r = \frac{1}{r!} \left(-\frac{1}{2}\right) \left(-\frac{3}{2}\right) \left(-\frac{5}{2}\right) \dots \left(-\frac{2r-1}{2}\right) = \frac{(-1)^r 1 \cdot 3 \cdot 5 \dots (2r-1)}{2^r r!}, \quad (44-56)$$

$$b_r = \frac{1}{r!} \left(-\frac{3}{2}\right) \left(-\frac{5}{2}\right) \left(-\frac{7}{2}\right) \dots \left(-\frac{2r+1}{2}\right) = \frac{(-1)^r 1 \cdot 3 \cdot 5 \dots (2r+1)}{2^r r!},$$

which can also be written in the form

$$a_r = (-1)^r \frac{(2r)!}{2^{2r} (r!)^2}, \quad b_r = (-1)^r \frac{(2r+1)!}{2^{2r} (r!)^2}. \quad (44-57)$$

The solution procedure may be described in the following way. First we calculate μ_n successively by (44-40), starting with (44-38). Then (44-54) and (44-55) give T_n , and T is finally obtained by (44-53) with $k = 1$, namely

$$T = \sum_{n=0}^{\infty} T_n. \quad (44-58)$$

The procedure is similar to Molodensky's solution as given by (43-32), (43-38), (43-40), (43-41). It is, however, simpler because there we had two sets of quantities, namely G_n and χ_n , whereas now we only have μ_n .

To get a better comparison with Molodensky's solution, let us write (44-40) and (44-55) explicitly for $n = 0, 1, 2, \dots$:

$$\begin{aligned}\mu_0 &= \Delta g, \\ \mu_1 &= \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_P}{l_0^3} \mu_0 d\sigma, \\ \mu_2 &= \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_P}{l_0^3} \mu_1 d\sigma + \mu_0 \tan^2 \beta, \\ \mu_3 &= \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_P}{l_0^3} \mu_2 d\sigma - \frac{3R^2}{4\pi} \iint_{\sigma} \frac{(h-h_P)^3}{l_0^5} \mu_0 d\sigma + \mu_1 \tan^2 \beta, \\ &\dots\end{aligned}\tag{44-59}$$

and

$$\begin{aligned}T_0 &= \frac{R}{4\pi} \iint_{\sigma} \mu_0 [S(\psi) - 1] d\sigma, \\ T_1 &= \frac{R}{4\pi} \iint_{\sigma} \mu_1 [S(\psi) - 1] d\sigma, \\ T_2 &= \frac{R}{4\pi} \iint_{\sigma} \mu_2 [S(\psi) - 1] d\sigma - \frac{R^2}{4\pi} \iint_{\sigma} \frac{(h-h_P)^2}{l_0^3} \mu_0 d\sigma, \\ T_3 &= \frac{R}{4\pi} \iint_{\sigma} \mu_3 [S(\psi) - 1] d\sigma - \frac{R^2}{4\pi} \iint_{\sigma} \frac{(h-h_P)^2}{l_0^3} \mu_1 d\sigma, \\ &\dots\end{aligned}\tag{44-60}$$

Here we have used numerical values for the coefficients a_r and b_r as given by (44-56).

The replacement of $S(\psi) - 1$ in these formulas by $S(\psi)$ changes T only by a constant

$$\frac{R}{4\pi} \iint_{\sigma} (\mu_0 + \mu_1 + \mu_2 + \dots) d\sigma,\tag{44-61}$$

which is frequently disregarded; it is zero if the mass of the reference ellipsoid equals the mass of the earth.

45. SOLUTION BY ANALYTICAL CONTINUATION

An elementary approach, avoiding integral equations, is possible through formal analytical continuation by means of a Taylor series. It extends the solution described in sec. 8-8 of (Heiskanen and Moritz, 1967) to higher-order approximations.

Continuation to point level. Let Δg be the gravity anomaly at the telluroid, and $\Delta g'$ be the free-air anomaly at point level, that is, on the normal level surface

$$U = U_A = \text{const.}, \quad (45-1)$$

A being the telluroid point at which the height anomaly ζ or the deflection of the vertical (ξ, η) is to be computed (Fig.45.1).

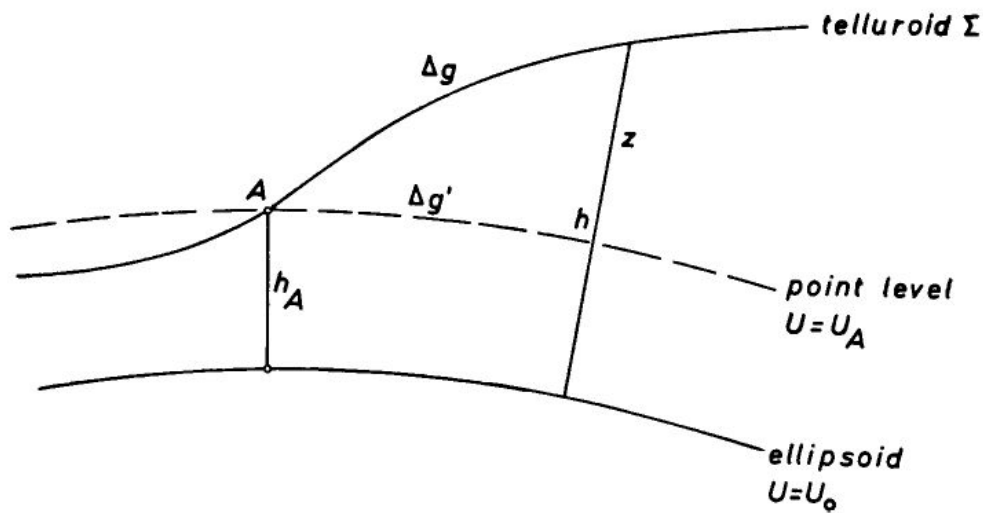


FIGURE 45.1. Point level.

On those parts of the point-level surface which are outside the telluroid (e.g., the part to the left of A in Fig.45.1), the gravity anomaly $\Delta g'$ corresponds to the external anomalous gravitational potential T : it is related to it by (43-2):

$$\Delta g = - \frac{\partial T}{\partial r} - \frac{2}{r} T ; \quad (45-2)$$

also in this section we use the spherical approximation. On those parts of the point-level surface which are below the telluroid, however, $\Delta g'$ corresponds to the *analytical continuation* T of the external potential T into the earth's interior. For the time being we assume that such an analytical continuation is possible; we shall return to this problem later.

Thus, T outside the telluroid Σ and T inside Σ together form a single harmonic function, which is assumed regular in the region needed (down to point level). Hence, also Δg and $\Delta g'$ are restrictions of the same analytic function: Δg to Σ and $\Delta g'$ to $U = U_A$. They are thus connected by a Taylor series:

$$\begin{aligned}\Delta g &= \Delta g' + z \frac{\partial \Delta g'}{\partial z} + \frac{1}{2!} z^2 \frac{\partial^2 \Delta g'}{\partial z^2} + \frac{1}{3!} z^3 \frac{\partial^3 \Delta g'}{\partial z^3} + \dots \\ &= \Delta g' + \sum_{n=1}^{\infty} \frac{1}{n!} z^n \frac{\partial^n \Delta g'}{\partial z^n},\end{aligned}\quad (45-3)$$

where

$$z = h - h_A \quad (45-4)$$

is the elevation difference with respect to the computation point A . For the present we shall assume the series (45-3) to be convergent.

Note that the derivatives $\partial/\partial z$, $\partial^2/\partial z^2$, ... in (45-3) designate derivatives with respect to the quantity (45-4); they are thus vertical derivatives. As a spherical approximation, they are radial derivatives:

$$\frac{\partial^n}{\partial z^n} = \frac{\partial^n}{\partial r^n}.$$

This series may be symbolically written as

$$\Delta g = U \Delta g', \quad (45-5)$$

where the symbol U denotes the *upward continuation operator*, which stands for the operation to be performed on the function $\Delta g'$ to get the function Δg according to (45-3).¹

¹The name, upward continuation operator, is not fully correct literally since the transition from $\Delta g'$ to Δg may also involve downward continuation (e.g., left of point A in Fig. 45.1). A similar remark holds for the name, downward continuation operator. It might be more correct to call U the "direct continuation operator" and D the "inverse continuation operator".

We are given Δg at the earth's surface. Let us compute $\Delta g'$ by some inversion of (45-3):

$$\Delta g' = U^{-1} \Delta g = D \Delta g, \quad (45-6)$$

where D , the downward continuation operator, is inverse to U .

Since $\Delta g'$ refers to a level surface, we may apply Stokes' formula (2-35) and Vening Meinesz' formula (2-40) to get the anomalous potential T and the deflection of the vertical (ξ, η) , all at the point A :

$$T = \frac{R}{4\pi} \iint_{\sigma} \Delta g' S(\psi) d\sigma, \quad (45-7)$$

$$\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \frac{1}{4\pi\gamma^0} \iint_{\sigma} \Delta g' \frac{dS}{d\psi} \begin{bmatrix} \cos \alpha \\ \sin \alpha \end{bmatrix} d\sigma. \quad (45-8)$$

Strictly speaking, we should in (45-7) replace R by $R + h_A$, but we may use R without impairing the accuracy; in the same way we may in (45-8) use a mean value γ^0 of about 980 gal as usual.

Derivation of $\Delta g'$. There remains now to compute the point level anomaly $\Delta g'$ from the measured ground anomaly Δg . We write (45-3) symbolically in the form

$$\begin{aligned} \Delta g &= \Delta g' + \left(\sum_{n=1}^{\infty} \frac{1}{n!} z^n \frac{\partial^n}{\partial z^n} \right) \Delta g' \\ &= \left(I + \sum_{n=1}^{\infty} z^n L_n \right) \Delta g'; \end{aligned} \quad (45-9)$$

$$L_n = \frac{1}{n!} \frac{\partial^n}{\partial z^n} = \frac{1}{n!} \frac{\partial^n}{\partial r^n} \quad (45-10)$$

is a vertical (radial) differentiation operator and I is the identity operator:

$$I f = f. \quad (45-11)$$

Comparing (45-9) with (45-5) we see that we have obtained a symbolic series expansion of the upward continuation operator U :

$$U = I + \sum_{n=1}^{\infty} z^n L_n . \quad (45-12)$$

We shall now try to compute the downward continuation operator

$$D = U^{-1} \quad (45-13)$$

by forming the formal reciprocal of the series (45-12); then (45-6) gives $\Delta g'$. This will be done as follows.

We replace all elevations h by kh , where k is the Molodensky parameter with $0 \leq k \leq 1$, as also used in the two preceding sections. Then the upward continuation operator (45-12) becomes

$$U = I + \sum_{n=1}^{\infty} k^n z^n L_n = \sum_{n=0}^{\infty} k^n U_n , \quad (45-14)$$

where

$$U_0 = I ; \quad U_n = z^n L_n \quad \text{if } n = 1, 2, 3, \dots \quad (45-15)$$

In the same way we express the downward continuation operator $D = U^{-1}$ as a formal series

$$D = \sum_{n=0}^{\infty} k^n D_n . \quad (45-16)$$

We may try to determine the D_n from the obvious operator identity

$$UD = I .$$

On substituting the respective series we have

$$\sum_{p=0}^{\infty} k^p U_p \sum_{q=0}^{\infty} k^q D_q = I$$

or

$$\sum_{p=0}^{\infty} \sum_{q=0}^{\infty} k^{p+q} U_p D_q = I .$$

The application of (44-33) gives

$$\sum_{n=0}^{\infty} k^n \sum_{r=0}^n U_r D_{n-r} - I = 0 .$$

We require this identity to hold for all values of the parameter k . Then the factors of all k^n must be zero. For $n = 0$ we have

$$U_0 D_0 - I = 0 ,$$

thus because of $U_0 = I$ also

$$D_0 = I .$$

For $n \neq 0$ we have the equation

$$\sum_{r=0}^n U_r D_{n-r} = 0 \quad (45-17)$$

or

$$D_n + \sum_{r=1}^n U_r D_{n-r} = 0 ,$$

whence

$$D_n = - \sum_{r=1}^n U_r D_{n-r} = 0 . \quad (45-18)$$

This equation expresses D_n in terms of the known U_r and the previously determined D_1, D_2, \dots, D_{n-1} . So, starting from $D_0 = I$, we can recursively compute the operators D_1, D_2, D_3, \dots .

Computationally more convenient is the introduction of the functions

$$g_n = D_n(\Delta g) . \quad (45-19)$$

Eq. (45-17) gives for them

$$\sum_{r=0}^n U_r D_{n-r}(\Delta g) = 0 .$$

By (45-15) and (45-19) this becomes

$$\sum_{r=0}^n z^r L_r(g_{n-r}) = 0, \quad (45-20)$$

which can be solved for g_n , noting $z^0 L_0(g_n) = g_n$,

$$g_n = - \sum_{r=1}^n z^r L_r(g_{n-r}). \quad (45-21)$$

Eq. (45-21) makes it possible to determine the g_n recursively, starting from

$$g_0 = \Delta g. \quad (45-22)$$

Then the anomaly $\Delta g'$, defined by (45-6), is then given by

$$\Delta g' = D\Delta g = \sum_{n=0}^{\infty} D_n(\Delta g) = \sum_{n=0}^{\infty} g_n. \quad (45-23)$$

We have put $k = 1$ in (45-16), so as to change kh back into the actual evaluation h , since we had admitted a general k only in order to get a convenient mechanism of expansion.

Then (45-7) gives

$$T = S(\Delta g') = \sum_{n=0}^{\infty} T_n \quad (45-24)$$

with

$$T_n = S(g_n), \quad (45-25)$$

S denoting the Stokes operator.

Determination of the L_n . We must now study the operators L_n which play a basic role in the present method.

First we derive some simple formulas for them. The definition (45-10) gives

$$L_n = \frac{1}{n!} \frac{\partial^n}{\partial z^n} = \frac{1}{n} \frac{1}{(n-1)!} \frac{\partial^{n-1}}{\partial z^{n-1}} \frac{\partial}{\partial z} \quad (45-26)$$

or

$$L_n = \frac{1}{n} L_{n-1} L = \frac{1}{n} L L_{n-1} . \quad (45-27)$$

This is a *recursion formula* expressing L_n in terms of L_{n-1} and $L = L_1$. Repeated application of this recursion formula gives

$$L_n = \frac{1}{n!} L^n \quad (45-28)$$

where

$$L^n = LLL \dots L \quad (n \text{ times}) ;$$

this is also evident from (45-26).

The original meaning of L_n as a spatial operator, namely a vertical derivative, is restricted to the use with level-surface anomalies $\Delta g'$ only; furthermore, this vertical derivative is normal to the surface and thus, figuratively speaking, leads out of the surface.

It is possible, however, to interpret L_n as a *surface operator* which does not lead out of the surface and can be used with data given on an arbitrary smooth surface which need not be a level surface. This may be done as follows.

The vertical derivative $\partial/\partial r$ can be expressed in terms of surface values by the well-known spherical formula (Heiskanen and Moritz, 1967, p. 38)

$$\frac{\partial f}{\partial r} = -\frac{1}{R} f + \frac{R^2}{2\pi} \iint_{\sigma} \frac{f - f_P}{l_O^3} d\sigma , \quad (45-29)$$

which again uses our current notational convention (p. 356), P being the point at which $\partial f/\partial r$ is computed and to which f in the first term on the right-hand side refers, too. Also the other notations are the same as before, σ denoting the unit sphere and

$$l_O = 2R \sin \frac{\psi}{2} . \quad (45-30)$$

As a planar approximation we may neglect the small term f/R in (45-29), so that the basic operator (45-26) becomes the surface operator

$$L(f) = \frac{R^2}{2\pi} \iint_{\sigma} \frac{f - f_P}{r_{PO}^3} d\sigma = L_1(f) . \quad (45-31)$$

The second derivative

$$\frac{\partial^2}{\partial r^2} = \frac{\partial^2}{\partial z^2} = 2L_2$$

can be expressed as a surface operator even more easily.

Let xyz be a local cartesian coordinate system at the point under consideration, the xy -plane being the tangent plane and the z -axis being vertical, in agreement with the notation (45-31). As a planar approximation, Δg is a harmonic function in space (p. 176), satisfying Laplace's equation

$$\frac{\partial^2 \Delta g}{\partial x^2} + \frac{\partial^2 \Delta g}{\partial y^2} + \frac{\partial^2 \Delta g}{\partial z^2} = 0 . \quad (45-32)$$

Therefore, (45-31) gives

$$L_2(\Delta g') = -\frac{1}{2} \left(\frac{\partial^2 \Delta g'}{\partial x^2} + \frac{\partial^2 \Delta g'}{\partial y^2} \right) . \quad (45-33)$$

Thus we may extend the definition of L_2 to arbitrary smooth surface functions:

$$L_2(f) = -\frac{1}{2} (f_{xx} + f_{yy}) , \quad (45-34)$$

the subscripts x and y denoting partial differentiation.

This equation is the planar approximation to the Laplacian surface operator Δ_2 for an arbitrary surface, as given, e.g., by Hotine (1969, p.45); Hotine denotes it by $\bar{\Delta}$, whereas we are using the more frequently employed symbol Δ_2 , where the subscript 2 expresses the two-dimensional character of a surface. For a sphere of radius R we have in geographical coordinates ϕ, λ :

$$\Delta_2 f = R^{-2} (f_{\phi\phi} + \cos^{-2} \phi f_{\lambda\lambda} - f_{\phi} \tan \phi) , \quad (45-35)$$

the subscripts ϕ and λ again denoting partial differentiation. Thus we may write (45-34) in the slightly more general form

$$L_2(f) = -\frac{1}{2} \Delta_2 f . \quad (45-36)$$

As a matter of fact, we could also express L_2 by (45-28), applying the operator L twice:

$$L_2(f) = \frac{1}{2} L^2(f) = \frac{1}{2} L[L(f)] ; \quad (45-37)$$

explicitly, we have by means of the auxiliary quantity

$$f_1 = L_1(f) = L(f) \quad (45-38)$$

the result

$$L_2(f) = \frac{1}{2} L(f_1) . \quad (45-39)$$

Generally we can by (45-27) express L_n recursively in terms of the surface operator (45-31): put

$$L_n(f) = f_{(n)} ; \quad (45-40)$$

then

$$\begin{aligned} f_{(1)} &= L(f) , \\ f_{(2)} &= \frac{1}{2} L(f_{(1)}) , \\ f_{(3)} &= \frac{1}{3} L(f_{(2)}) , \\ &\vdots \\ f_{(n)} &= \frac{1}{n} L(f_{(n-1)}) . \end{aligned} \quad (45-41)$$

This definition of the L_n as surface operators--by (45-31), (45-36), and the recursion formula (45-27)--is the relevant one for the present purpose. For instance, $L_n(\Delta g)$ is to be understood in this way; it would be wrong to interpret it as

$$\frac{1}{n!} \frac{\partial^n \Delta g}{\partial r^n} ,$$

as a vertical derivative at the telluroid Σ , since Δg at Σ does not refer to a level surface.

Computational formulas. Let us finally summarize our computational formulas. By Bruns' equation (42-16) and by (45-7), (45-8), and (45-23) we have

$$\zeta = \frac{R}{4\pi\gamma^0} \iint_{\sigma} \Delta g S(\psi) d\sigma + \sum_{n=1}^{\infty} \frac{R}{4\pi\gamma^0} \iint_{\sigma} g_n S(\psi) d\sigma , \quad (45-42)$$

$$\begin{Bmatrix} \xi \\ \eta \end{Bmatrix} = \frac{1}{4\pi\gamma^0} \iint_{\sigma} \Delta g \frac{dS}{d\psi} \begin{Bmatrix} \cos \alpha \\ \sin \alpha \end{Bmatrix} d\sigma + \sum_{n=1}^{\infty} \frac{1}{4\pi\gamma^0} \iint_{\sigma} g_n \frac{dS}{d\psi} \begin{Bmatrix} \cos \alpha \\ \sin \alpha \end{Bmatrix} d\sigma . \quad (45-43)$$

Here γ^0 is a global mean value such as 980 gal. The correction terms g_n are evaluated recursively by (45-21):

$$g_n = - \sum_{r=1}^n z^r L_r(g_{n-r}) , \quad (45-44)$$

starting from

$$g_0 = \Delta g ; \quad (45-45)$$

there is

$$z = h - h_A . \quad (45-46)$$

The L_n are also evaluated recursively:

$$L_n(\Delta g) = \frac{1}{n} L_1 [L_{n-1}(\Delta g)] \quad (45-47)$$

with

$$L_1(f) = \frac{R^2}{2\pi} \iint_{\sigma} \frac{f - f_p}{r_{op}^3} d\sigma \quad (45-48)$$

These formulas are all that is needed to compute approximations of an arbitrarily high order. All occurring operators are systematically reduced to a repeated application of the integral (45-48).

Let us finally render the method more concrete by evaluating (45-44) explicitly for $n = 1, 2, 3$:

$$\begin{aligned} g_1 &= -z L_1(\Delta g) , \\ g_2 &= -z L_1(g_1) - z^2 L_2(\Delta g) , \\ g_3 &= -z L_1(g_2) - z^2 L_2(g_1) - z^3 L_3(\Delta g) . \end{aligned} \quad (45-49)$$

If we restrict ourselves to $n = 1$, then the present solution becomes

$$\zeta = \frac{R}{4\pi\gamma_0} \iint_{\sigma} \left[\Delta g - (h-h_A) \frac{\partial \Delta g}{\partial h} \right] S(\psi) d\sigma \quad (45-50)$$

since

$$g_1 = - (h-h_A) L_1(\Delta g) = - (h-h_A) \frac{\partial \Delta g}{\partial z} = - (h-h_A) \frac{\partial \Delta g}{\partial h} \quad (45-51)$$

and the operator L_1 may be interpreted as a vertical derivative by (45-10). This first-order solution may, therefore, be called *gradient solution*. Analogous formulas hold for ξ and η . All these formulas are very suitable for practical application; cf. (Heiskanen and Moritz, 1967, secs. 8-8 and 8-9).

The use of analytical continuation for the solution of Molodensky's problem has an interesting history which has its ups and downs. It was considered already by Molodensky in 1949 for a practical solution of his problem but later rejected because the required downward continuation cannot be expected to be regular (see below). Later Bjerhammar (1964) took up the idea and developed it in the way described in sec. 8-10 of (Heiskanen and Moritz, 1967). The first-order approximation (to $n = 1$) of the present method is given in secs. 8-8 and 8-9 of that book; it is also shown there that the solution is equivalent to Molodensky's series to $n = 1$.

The full series solution, as an expansion in terms of Molodensky's parameter k , was developed simultaneously and independently by Marych (1969) and Moritz (1969b). The latter used the present systematic approach in terms of an expansion of the operators U and D as formal power series with respect to k and gave general recursion formulas. He also showed the equivalence with Molodensky's series by an indirect argument using the theory of asymptotic series. The concise recursion formula (45-44) is due to Ecker (1971).

The preceding developments were effected in a purely formal manner, without regard to convergence problems and irrespective of the question whether the presupposed analytical continuation is possible at all. In fact, this analytical continuability cannot in general be assumed; cf. (Heiskanen and Moritz, 1967, p.321) and sec. 7 of the present book.

Thus a direct investigation of the mathematical validity and the convergence of the solution is difficult. Therefore we shall bypass these difficulties by exhibiting the termwise identity of the present solution to Molodensky's series in the following section; the convergence problem will then be considered in sec. 47.

Why, then, does the present solution lead to results that are as good as the approach by way of integral equations, although the analytical continuation of the external potential into the earth's interior cannot be expected to be regular? An intuitive explanation can be found in Runge's theorem (sec.8): even if the original potential T cannot be regularly continued down to the geoid (say), we can always find another harmonic function T' , arbitrarily close to T , which can be so continued and for which the basic Taylor expansion (45-3) converges.

46. PELLINEN'S EQUIVALENCE PROOF

In the three preceding sections we have met with three different expansions for the anomalous potential T . All three are formal series of powers of Molodensky's parameter k , but the first two--Molodensky's and Brovar's solutions--are obtained by a solution of integral equations, whereas the third expansion is based on a completely different principle, namely on analytical continuation by means of Taylor series. There arises the question how these different expansions are related among themselves.

An indirect argument demonstrates that all these three series expansions must be termwise equal. In fact, it can be shown that each of these three series, of form

$$T = \sum_{n=0}^{\infty} k^n T_n, \quad (46-1)$$

is an asymptotic series for $k \rightarrow 0$. Using the theory of asymptotic series we can conclude that any two asymptotic expansions of the same function T with respect to the same parameter k must be identical (Moritz, 1969b; 1971, sec.3).

However, a direct equivalence proof is highly desirable because it gives immediate insight into the structure of the various terms. For $n = 1$ this equivalence has been discussed already in sec. 8-8 of (Heiskanen and Moritz, 1967). For $n = 2$ the equivalence is verified in (Moritz, 1971, sec.4). These considerations have been extended to $n = 3$ by Ecker (1971), who also pointed out some lines along which a general equivalence proof could be achieved, and gave corresponding formulas; but there remained difficulties which looked forbidding.

These difficulties were resolved by Pellinen (1972), in particular, by finding eq. (46-9) below. We shall here present a slightly modified version of this important and elegant work.

Pellinen's identity. In the sequel, F, G, U, V, W, \dots denote functions of two variables (x, y) which are differentiable as often as required; x and y are rectangular coordinates in the plane. We introduce the plane Laplacian operator

$$\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \quad (46-2)$$

For the Laplacian of the product FG we find by direct differentiation:

$$\Delta(FG) = F\Delta G + G\Delta F + 2D(F, G) \quad (46-3)$$

where

$$D(F, G) = \frac{\partial F}{\partial x} \frac{\partial G}{\partial x} + \frac{\partial F}{\partial y} \frac{\partial G}{\partial y} \quad (46-4)$$

Another identity is

$$UD(V, V) = \frac{1}{2} \Delta(UV^2) - V\Delta(UV) + \frac{1}{2} V^2 \Delta U \quad (46-5)$$

This is verified by expressing the first two terms on the right-hand side by means of (46-3); then most terms on the right-hand side cancel and the left-hand side remains.

Using our basic notational convention (p.356), we may write (46-5) in the simplified form

$$UD(V,V) = \frac{1}{2} \Delta [(V-V_P)^2 U] . \quad (46-6)$$

This is directly seen by applying Δ to

$$\frac{1}{2}(V-V_P)^2 U = \frac{1}{2} UV^2 - V_P UV + \frac{1}{2} V_P^2 U ; \quad (46-7)$$

we have, e.g.,

$$\Delta(V_P UV) = V_P \Delta(UV) = V \Delta(UV) \quad (46-8)$$

since V_P is not affected by the operator Δ and becomes V outside of it, in agreement with our notational convention.

Eq. (46-6) and its generalization by Pellinen to higher $r = 1, 2, 3, \dots$,

$$U[D(V,V)]^r = \frac{1}{(2r)!} \Delta^r [(V-V_P)^{2r} U] , \quad (46-9)$$

will play a fundamental role. This formula will be called *Pellinen's identity*; it may be proved as follows.

By direct differentiation we find

$$\frac{\partial^k (W^k U)}{\partial x^{k-j} \partial y^j} = k! W_x^{k-j} W_y^j U + W[\dots] , \quad (46-10)$$

where W_x^{k-j} denotes the partial derivative W_x raised to the $(k-j)$ th power. The terms between brackets, which are multiplied by W , will not be needed later. As an example,

$$(W^2 U)_{xx} = 2W_x^2 U + 2WW_{xx} U + 4W W_x U_x + W^2 U_{xx} = 2! W_x^2 U + W[\dots] .$$

Put now

$$W = V - V_p ; \quad (46-11)$$

then

$$W_x = V_x \quad (46-12)$$

since V_p is constant with respect to differentiation. Outside the differentiation we have

$$V - V_p = 0 \quad (46-13)$$

since $V = V_p$ there, by our notational convention; cf. (46-8). Hence (46-10) gives

$$\frac{\partial^k [(V - V_p)^k U]}{\partial x^{k-j} \partial y^j} = k! V_x^{k-j} V_y^j U ; \quad (46-14)$$

the last term in (46-10) contains (46-13) as a factor and hence vanishes.

Consider now the r -th power of the Laplacian operator (46-3). The binomial theorem gives

$$\Delta^r = \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right)^r = \sum_{s=0}^r \binom{r}{s} \frac{\partial^{2r}}{\partial x^{2r-2s} \partial y^{2s}} , \quad (46-15)$$

so that

$$\Delta^r [(V - V_p)^{2r} U] = \sum_{s=0}^r \binom{r}{s} \frac{\partial^{2r} [(V - V_p)^{2r} U]}{\partial x^{2r-2s} \partial y^{2s}} .$$

By (46-14) with $k = 2r$ and $j = 2s$ this is equal to

$$(2r)! \sum_{s=0}^r \binom{r}{s} V_x^{2r-2s} V_y^{2s} U = (2r)! \left(V_x^2 + V_y^2 \right)^r U = (2r)! U [D(V, V)]^r ,$$

which was to be proved.

Other auxiliary formulas. The operator (45-31) becomes for a plane reference surface

$$L(f) = \frac{1}{2\pi} \iint \frac{f - f_P}{l^3} dS, \quad (46-16)$$

where the integral is extended over the whole plane:

$$\iint = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty}, \quad (46-17)$$

the surface element dS is given by

$$dS = dx dy, \quad (46-18)$$

and l replaces our former l_0 ; there is

$$l = \sqrt{(x - x_P)^2 + (y - y_P)^2}, \quad (46-19)$$

the point $P(x_P, y_P)$ being the point for which $L(f)$ is to be computed.

For the integral (46-16) to exist we must require that the function $f(x, y)$ tends to zero at infinity in a sufficiently rapid manner. This will always be assumed.

The second power of this operator L is, by (45-36) and (45-37), nothing else but the plane Laplacian:

$$L^2 = -\Delta; \quad (46-20)$$

we are now writing Δ instead of Δ_2 . We also recall the definition of L_n by (45-26):

$$L_n = \frac{1}{n!} L^n. \quad (46-21)$$

In view of (46-20), the Pellinen identity (46-9) may be expressed in terms of the $2r$ -th power of the operator L :

$$UD^r(V, V) = \frac{(-1)^r}{(2r)!} L^{2r}[(V - V_P)^{2r}U]; \quad (46-22)$$

we also have slightly simplified the notation on the left-hand side.

Stokes' formula takes for the plane the form

$$S(f) = \frac{1}{2\pi} \iint \frac{f}{T} dS . \quad (46-23)$$

To see this, write

$$T = \frac{R}{4\pi} \iint_{\sigma} \Delta g S(\psi) d\sigma = S(\Delta g) \quad (46-24)$$

as

$$S(\Delta g) = \frac{1}{4\pi} \iint_{\sigma} \Delta g R^{-1} S(\psi) dS , \quad (46-25)$$

where

$$dS = R^2 d\sigma \quad (46-26)$$

is now the surface element of the sphere $r = R$. (The use of the letter S both in Stokes' operator and in the surface element is accidental.) The function $R^{-1}S(\psi)$ is given by (44-43), and we get from (44-11):

$$R^{-1}S(\psi) = \frac{2}{1} + \frac{1}{R} [\dots] . \quad (46-27)$$

For $R \rightarrow \infty$ there remains $2/1$ where 1 is now given by the plane formula (46-19), and (46-25) becomes the integral

$$S(\Delta g) = \frac{1}{2\pi} \iint \frac{\Delta g}{1} dS , \quad (46-28)$$

extended over the plane, dS now being given by (46-18). This proves (46-23). The same planar approximation holds for (43-42).

It is not difficult to see that the plane operators L and S are related by

$$S = -L^{-1} . \quad (46-29)$$

In fact, the operator L gives the vertical derivative of f :

$$L(f) = \frac{\partial f}{\partial z} .$$

For the plane we have

$$\Delta g = - \frac{\partial T}{\partial z} = - LT = - LS(\Delta g) ,$$

using (46-24). Thus

$$LS = - I ,$$

(46-30)

I denoting the unit operator; but this is equivalent to (46-29). We finally recall the well-known *Green identity* for the plane:

$$\iint_S (F \Delta G - G \Delta F) dS = \int_C \left(F \frac{\partial G}{\partial n} - G \frac{\partial F}{\partial n} \right) dC ,$$

(46-31)

where C is a closed curve bounding a simply connected region S in the plane; $\partial/\partial n$ denotes the derivative normal to the curve C (positive outward), and dC is the arc element of the curve. This formula is the precise analogue for the plane of Green's second identity for three-dimensional space; cf. (Heiskanen and Moritz, 1967, p.11).

Consider now the curve C to be a circle of very large radius ρ , and let $\rho \rightarrow \infty$. If the functions F and G , together with their first derivatives, vanish sufficiently rapidly at infinity, then the integral on the right tends to zero, and there remains

$$\iint (F \Delta G - G \Delta F) dS = 0 ,$$

(46-32)

where the integral is extended over the whole plane in agreement with (46-17). This is the form in which we shall need Green's identity.

Ecker's formula. Consider the integral

$$J = \frac{1}{2\pi} \iint 1^{-(2r+1)} (V - V_P)^{2r} U dS ,$$

(46-33)

extended as usual over the plane; r is an integer ≥ 1 . Let us apply Green's identity to this integral.

We have

$$\Delta(1^{-n}) = n^2 1^{-(n+2)} ,$$

(46-34)

as one verifies by direct differentiation, using (46-19); hence

$$1^{-(2r+1)} = \frac{1}{(2r-1)^2} \Delta 1^{-(2r-1)} . \quad (46-35)$$

Thus the application of (46-32) with

$$F = (V - V_P)^{2r} U , \quad G = 1^{-(2r-1)} \quad (46-36)$$

gives

$$\begin{aligned} J &= \frac{1}{2\pi(2r-1)^2} \iint (V - V_P)^{2r} U \Delta 1^{-(2r-1)} dS \\ &= \frac{1}{2\pi(2r-1)^2} \iint 1^{-(2r-1)} \Delta [(V - V_P)^{2r} U] dS . \end{aligned} \quad (46-37)$$

The subsequent application of (46-32) with

$$F = \Delta [(V - V_P)^{2r} U] , \quad G = 1^{-(2r-3)} ,$$

noting that, by (46-34),

$$1^{-(2r-1)} = \frac{1}{(2r-3)^2} \Delta 1^{-(2r-3)} ,$$

yields

$$J = \frac{1}{2\pi(2r-1)^2(2r-3)^2} \iint 1^{-(2r-3)} \Delta^2 [(V - V_P)^{2r} U] dS .$$

Continuing in this way, we obtain¹

$$J = \frac{1}{2\pi(2r-1)^2(2r-3)^2 \dots 3^2} \iint 1^{-3} \Delta^{r-1} [(V - V_P)^{2r} U] dS . \quad (46-38)$$

We finally note that, by (46-20),

$$\Delta^{r-1} = (-1)^{r-1} L^{2r-2} ,$$

¹We remark that the application of the Green identity (46-32) to the function $1^{-(2r-1)}$ would have to be justified for full mathematical rigor, because of the singularity of this function.

and that we can express the integral in terms of the operator (46-16), putting

$$f = \Delta^{r-1} [(V - V_P)^{2r} U] ;$$

there is $f_P = 0$ since it contains $(V - V_P) = 0$ as factor even after performing the differentiations implied by Δ^{n-1} .

Thus there results

$$\frac{1}{2\pi} \iint \frac{(V - V_P)^{2r} U}{1^{2r+1}} dS = (-1)^{r-1} \left[\frac{2^r r!}{(2r)!} \right]^2 L^{2r-1} [(V - V_P)^{2r} U] ; \quad (46-39)$$

we have expressed the numerical factor in front of (46-38) in terms of factorials, just as we did in going from (44-56) to (44-57).

In exactly the same way we find

$$\frac{1}{2\pi} \iint \frac{(V - V_P)^{2r+1} U}{1^{2r+3}} dS = (-1)^r \left[\frac{2^r r!}{(2r+1)!} \right]^2 L^{2r+1} [(V - V_P)^{2r+1} U] . \quad (46-40)$$

These two formulas were derived by Ecker (1971).

Transformation of Brovar's formula. After these lengthy preparations, the equivalence proof is rather straightforward. We write Brovar's solution, as given by (44-40) and (44-55), for a plane reference surface, which we use for $n \geq 1$. (As a matter of fact, Stokes' approximation, which corresponds to $n = 0$, remains spherical.) We have for $n \geq 1$:

$$T_n = S(\mu_n) + \sum_{r=1}^M (-1)^r \frac{(2r)!}{2^{2r} (r!)^2} \frac{1}{2\pi} \iint \frac{(h - h_P)^{2r}}{1^{2r+1}} \mu_{n-2r} dS , \quad (46-41)$$

$$\begin{aligned} \mu_n &= \sum_{r=0}^N (-1)^r \frac{(2r+1)!}{2^{2r} (r!)^2} \frac{1}{2\pi} \iint \frac{(h - h_P)^{2r+1}}{1^{2r+3}} \mu_{n-2r-1} dS - \\ &\quad - \sum_{r=1}^M (-1)^r D^r(h, h) \mu_{n-2r} . \end{aligned} \quad (46-42)$$

Here we have expressed the coefficients a_r and b_r by (44-57) and $\tan^{2r} \beta$ by

$$\tan^2 \beta = h_x^2 + h_y^2 = D(h, h) , \quad (46-43)$$

which relates the maximum inclination β of the terrain to the elevation h in a geometrically evident way. For $n = 0$ there is simply

$$\mu_0 = \Delta g. \quad (46-44)$$

The upper limits of the sums, M and N , are given by (44-37).

The transformation of (46-41) by means of Ecker's formula (46-39), with $U = \mu_{n-2r}$ and $V = h$, gives

$$T_n = S(\mu_n) - \sum_{r=1}^M \frac{1}{(2r)!} L^{2r-1} \left[(h - h_A)^{2r} \mu_{n-2r} \right], \quad (46-45)$$

where we have denoted by A the point at which T is computed. Then (46-42) is transformed using Pellinen's identity (46-22) and Ecker's formula (46-40):

$$\begin{aligned} \mu_n = & \sum_{r=0}^N \frac{1}{(2r+1)!} L^{2r+1} \left[(h - h_P)^{2r+1} \mu_{n-2r-1} \right] \\ & - \sum_{r=1}^M \frac{1}{(2r)!} L^{2r} \left[(h - h_P)^{2r} \mu_{n-2r} \right]. \end{aligned} \quad (46-46)$$

This reduces simply to

$$\mu_n = - \sum_{k=1}^n \frac{(-1)^k}{k!} L^k \left[(h - h_P)^k \mu_{n-k} \right], \quad (46-47)$$

since the first sum in (46-46) gives the terms with odd $m = 2r+1$ and the second sum provides the terms with even $m = 2r$. The summation variable k has, of course, nothing in common with Molodensky's shrinking parameter.

A last simplification is achieved by introducing the operator (45-28),

$$L_n = \frac{1}{n!} L^n, \quad (46-48)$$

so that (46-47) becomes

$$\mu_n = - \sum_{k=1}^n (-1)^k L_k \left[(h - h_P)^k \mu_{n-k} \right]. \quad (46-49)$$

Derivation of the analytical continuation solution from Brovar's series. Finally we shall derive the analytical continuation solution from the transformed Brovar formulas (46-45) and (46-47). Let us introduce the elevation above the level of the computation point¹ A by

$$z = h - h_A ; \quad (46-50)$$

this is in agreement with (45-4).

In (46-49) we combine both sides of the equation into one sum $\sum_{k=0}^n$ since $L_0 = I$, and expand

$$(h - h_P)^k = (z - z_P)^k = \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} z_P^{k-j} z^j \quad (46-51)$$

by the binomial theorem. Since $(-1)^{2k-j} = (-1)^j$ we obtain

$$\sum_{k=0}^n \sum_{j=0}^k (-1)^j \binom{k}{j} z^{k-j} L_k \left(z^j \mu_{n-k} \right) = 0 ; \quad (46-52)$$

note that z_P becomes z outside the operator L_k . Now we introduce new summation indices r, s by putting

$$r = k - j, \quad s = j, \quad \text{whence} \quad k = r + s, \quad (46-53)$$

which results in

$$\sum_{r=0}^n \sum_{s=0}^{n-r} (-1)^s \binom{r+s}{s} z^r L_{r+s} \left(z^s \mu_{n-r-s} \right) = 0 . \quad (46-54)$$

But by (46-48) we have

$$L_{r+s} = \frac{1}{(r+s)!} L^{r+s} = \frac{r!s!}{(r+s)!} L_r L_s = \binom{r+s}{s}^{-1} L_r L_s, \quad (46-55)$$

according to the definition of the binomial coefficients. Thus (46-54) reduces to

¹ A denotes the computation point for T , whereas the point at which μ is computed continues to be designated by P .

$$\sum_{r=0}^n z^r L_r \left[\sum_{s=0}^{n-r} (-1)^s L_s \left(z^s \mu_{n-r-s} \right) \right] = 0 . \quad (46-56)$$

On introducing the quantities

$$g_m = \sum_{s=0}^m (-1)^s L_s \left(z^s \mu_{m-s} \right) \quad (46-57)$$

this becomes

$$\sum_{r=0}^n z^r L_r (g_{n-r}) = 0 . \quad (46-58)$$

However, this is precisely the equation (45-20) by which the quantities (45-19) are uniquely defined, starting from

$$g_0 = \Delta g .$$

Since (46-57) gives for $m = 0$:

$$g_0 = \mu_0 = \Delta g ;$$

there is agreement also for the terms with $n = 0$. From this we conclude that the quantities (46-57) are identical to the quantities (45-19) denoted by the same symbol in the solution of sec. 45.

By (46-57) for $m = n$ we have

$$\mu_n = g_n - \sum_{s=1}^n (-1)^s L_s \left(z^s \mu_{n-s} \right) . \quad (46-59)$$

The substitution into (46-45) gives

$$\begin{aligned} T_n = S(g_n) - \sum_{s=1}^n (-1)^s S L_s \left(z^s \mu_{n-s} \right) \\ - \sum_{r=1}^M \frac{1}{(2r)!} L^{2r-1} \left(z^{2r} \mu_{n-2r} \right) , \end{aligned} \quad (46-60)$$

in view of (46-50). The first sum on the right-hand side becomes, on using (46-48) and (46-30),

$$\sum_{s=1}^n (-1)^s \frac{1}{s!} L^{s-1} \left(z^s u_{n-s} \right) .$$

We split up this sum into two parts, one consisting of the terms with even $s = 2r$ and the other consisting of the terms with odd $s = 2r+1$. The first part cancels with the second sum in (46-60), and there remains

$$- \sum_{r=0}^N \frac{1}{(2r+1)!} L^{2r} \left(z^{2r+1} u_{n-2r-1} \right) . \quad (46-61)$$

Now L^{2r} , apart from the sign, is Δ^r , which is a differentiation operator of order $2r$. We readily see that, after performing the differentiation, all terms contain z as a factor. Now, for the computation point A ,

$$z = z_A = 0$$

by (46-50). Therefore all terms contain zero as a factor, so that (46-61) vanishes.

Hence (46-60) reduces to

$$T_n = S(g_n) , \quad (46-62)$$

in agreement with (45-25). Thus we have derived the solution by analytical continuation from Brovar's solution, which completes the proof of equivalence of the two solutions.

In order to get a mathematically "clean" situation and to avoid the need of ad hoc neglecting terms during the course of the derivation, we have separated the zero-order terms, corresponding to $n = 0$ and given simply by Stokes' formula, from the terms with $n \geq 1$. For the zero-order terms we have used the sphere as a reference surface; for the higher-order terms, the reference surface is a plane. (In this way it is possible to give a precise mathematical definition of the "spherical" and the "planar" approximation, although such a consequent interpretation is not, in general, practically desirable; in operational formulas and derivations it is usually preferable to regard the planar approximation as the omission of terms of order h/R , still maintaining the sphere as a reference surface.)

This formal separation of the Stokes' term with $n = 0$ from the higher terms is possible because, for $h = 0$, even Brovar's rigorous spherical integral equation (44-22), without planar approximation, gives $u_0 = \Delta g$; likewise, for the solution of sec. 45, we get $g_0 = \Delta g$ spherically as well as in the planar case.

It is clear that, in order to consequently carry through the concept of a reference plane, we have to assume Δg as a smooth (infinitely differentiable) function given in the plane and vanishing appropriately at infinity.

Molodensky's and Brovar's series. Compared with the preceding proof of equivalence between Brovar's series and the series solution obtained by analytical continuation, it is almost trivial to see the equivalence between Molodensky's series, as discussed in sec. 43, and Brovar's series treated in sec. 44. If the reference surface is considered a plane, then Stokes' function reduces to $2/1$, so that Molodensky's representation (43-3) and Brovar's representation (44-13) become equivalent, and there is

$$2\pi\chi_n = G_n = u_n \quad (n \geq 1) \quad (46-63)$$

for a plane reference surface.

So, as a planar approximation, the correction terms in all three solutions are equal. The analytical continuation solution is as valid, or as questionable, as Molodensky's series. A study of the convergence behavior of any of these three--essentially identical--series is, therefore, in order; this will be attempted in the next section.

47, CONVERGENCE OF MOLODENSKY'S SERIES

In view of the equivalence of various series solutions of Molodensky's problem, as discussed in the preceding section, we can limit our considerations to one of these series. We select Brovar's solution (sec.44), which offers the simplest approach to the convergence problem. We shall here only present the main points of the proof; more details will be found in the report (Moritz, 1972), of which (Moritz, 1973b) is a shortened version.

Brovar's integral equation and its iterative solution. We start with Brovar's integral equation in the form (44-23)¹ and with the same notations:

¹ For the rigorous spherical integral equation (44-22) we should get the same result concerning convergence.

$$u \cos^2 \beta - \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_P}{l^3} u d\sigma = \Delta g. \quad (47-1)$$

On introducing a new auxiliary function

$$\kappa = u \cos^2 \beta, \quad (47-2)$$

this integral equation becomes

$$\kappa - \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_P}{l^3 \cos^2 \beta} \kappa d\sigma = \Delta g. \quad (47-3)$$

This equation may be written symbolically as

$$(I - \Phi)\kappa = \Delta g, \quad (47-4)$$

where I is the identity operator, with $If = f$, and Φ is the integral operator given by

$$\Phi f = \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_P}{l^3 \cos^2 \beta} f d\sigma \quad (47-5)$$

for an arbitrary function for which the integral exists.

The integral (47-5) is *strongly singular* since the kernel $(h-h_P)/l^3 \cos^2 \beta$ becomes infinite as $1/l^2$ if $l \rightarrow 0$ (note that $(h-h_P)/l$ remains bounded as $l \rightarrow 0$). The integral must be understood in the sense of the *Cauchy principal value*, that is, a small circle $\psi \leq \epsilon$ around the computation point P is first excluded from the unit sphere and the integral is then defined as

$$\lim_{\epsilon \rightarrow 0} \iint_{\psi > \epsilon} \quad . \quad (47-6)$$

The integral (47-5) would not exist as an ordinary improper integral, as "weakly singular" integrals do. (An example of a weakly singular integral is Stokes' integral in which the kernel becomes infinite only as $1/l$.)

The classical boundary-value problems of Dirichlet and Neumann lead to weakly singular integral equations, which are of Fredholm type; cf. (Kellogg, 1929, chapter XI). Contrary to this, Molodensky's problem, as an

oblique derivative problem, leads to strongly singular integral equations; cf. (Miranda, 1970, sec.23). These are not of Fredholm type and are not considered in standard textbooks on potential theory. The first book on strongly singular integral equations in two and more dimensions seems to be (Mikhlin, 1965); this book is still very useful for the present purpose. A formal solution of (47-3) is obtained by writing

$$\kappa = (I - \Phi)^{-1} \Delta g \quad (47-7)$$

with the Neumann series

$$(I - \Phi)^{-1} = I + \Phi + \Phi^2 + \Phi^3 + \dots \quad (47-8)$$

Such series are known from linear algebra, where I and Φ denote matrices; they are generalizations of the elementary binomial series

$$(1 - x)^{-1} = 1 + x + x^2 + x^3 + \dots \quad (47-9)$$

Here Φ^2 means, of course,

$$\Phi^2 f = \Phi(\Phi f), \quad (47-10)$$

and similarly for higher powers.

The series (47-9) converges for $|x| < 1$. In full analogy, the series (47-8) may be shown to converge in a normed space of operators if

$$\|\Phi\| < 1, \quad (47-11)$$

$\|\Phi\|$ denoting the norm of the operator Φ ; cf. sec. 5.

Equations (47-7) and (47-8) give the simplest solution of Brovar's integral equation from the point of view of functional analysis (provided Neumann's series converges!). This solution is different from a series of Molodensky type but essentially identical to the iterative solution described by (44-24). In fact, applying this iterative procedure to the operator equation (47-4), we get

$$\kappa^{(1)} = \Delta g,$$

$$\kappa^{(2)} = \Delta g + \Phi \kappa^{(1)} = (I + \Phi) \Delta g,$$

$$\kappa^{(3)} = \Delta g + \phi \kappa^{(2)} = (I + \phi + \phi^2) \Delta g , \quad (47-12)$$

. . . .

$$\kappa^{(n)} = \Delta g + \phi \kappa^{(n-1)} = (I + \phi + \phi^2 + \dots + \phi^{n-1}) \Delta g ,$$

which, on letting $n \rightarrow \infty$, is seen to give the Neumann series solution.

Unfortunately the convergence condition (47-11) cannot be guaranteed to hold, as we shall see later in this section.

Introduction of Molodensky's parameter. Let us now introduce Molodensky's shrinking parameter k , which has played an essential role in the preceding sections. Thus we replace all elevations h by kh ($0 \leq k \leq 1$).

Then the operator ϕ as given by (47-5) is replaced by $k\phi_k$ where ϕ_k is defined by

$$\phi_k f = \frac{R^2}{2\pi} \iint_{\sigma} \frac{h-h_P}{l_k^3 \cos^2 \beta_k} f d\sigma , \quad (47-13)$$

with

$$\tan \beta_k = k \tan \beta , \quad (47-14)$$

$$l_k^2 = l_0^2 \left[1 + k^2 \left(\frac{h-h_P}{l_0} \right)^2 \right] , \quad (47-15)$$

$$l_0 = 2R \sin \frac{\psi}{2} \quad (47-16)$$

as usual (sec.43).

In the place of (47-4) we now have the operator equation

$$(I - k\phi_k)\kappa = \Delta g , \quad (47-17)$$

which for $k = 1$ reduces to (47-4). A formal solution is given by the Neumann series

$$\kappa = (I + k\phi_k + k^2\phi_k^2 + k^3\phi_k^3 + \dots) \Delta g , \quad (47-18)$$

which converges for

$$k \|\phi_k\| < 1 , \quad (47-19)$$

according to (47-11).

Let now all operators Φ_k be uniformly bounded for $0 \leq k \leq 1$, that is,

$$\sup_{0 \leq k \leq 1} \|\Phi_k\| = C < \infty. \quad (47-20)$$

Then the Neumann series (47-18) will converge in the interval $0 \leq k < k_0$ with

$$k_0 = C^{-1} \quad (47-21)$$

as (47-19) is then satisfied. The gist of the lengthy argument to follow is that *Molodensky's series converges for the same values of k* , so that the convergence of Molodensky's series is determined by the convergence of Neumann's series (47-18).

Molodensky's series. The reason why (47-18) is not yet of Molodensky type is that the operator Φ_k and its powers also depend on k . Let us now expand the operators $\Phi_k, \Phi_k^2, \Phi_k^3, \dots$ themselves into series of powers of k , substitute these series into (47-18), and collect terms multiplied by equal powers of k . We thus get a series

$$\kappa = \sum_{n=0}^{\infty} k^n \kappa_n, \quad (47-22)$$

of which the functions κ_n are now independent of κ ; there is, for instance,

$$\kappa_0 = \Delta g.$$

The expansion (47-22) constitutes the Molodensky series for the present problem.

The convergence of (47-22) will be understood in the sense of convergence in norm:

$$\lim_{N \rightarrow \infty} \left\| \kappa - \sum_{n=0}^N k^n \kappa_n \right\| = 0. \quad (47-23)$$

A sufficient condition for convergence is *absolute convergence* (Dieudonné, 1960, sec.5.3):

$$\sum_{n=0}^{\infty} k^n \| \kappa_n \| < \infty ; \quad (47-24)$$

we shall work with absolutely convergent series.

The operators Φ_k are given by (47-13). If we expand $1/l_k^3$ by (44-28) and put

$$\cos^{-2} \beta_k = 1 + k^2 \tan^2 \beta , \quad (47-25)$$

then we obtain a series expansion of the form

$$\Phi_k = \sum_{r=0}^{\infty} k^{2r} B_{2r} , \quad (47-26)$$

where the operators B_{2r} are given by

$$B_0 f = \frac{R^2}{2\pi} \iint_{\sigma} f l_0^{-2} n d\sigma , \quad (47-27)$$

$$\begin{aligned} B_{2r} f = & b_r \frac{R^2}{2\pi} \iint_{\sigma} f l_0^{-2} n^{2r+1} d\sigma + \\ & + b_{r-1} \frac{R^2}{2\pi} \iint_{\sigma} f l_0^{-2} n^{2r-1} \tan^2 \beta d\sigma \quad (r > 0) . \end{aligned} \quad (47-28)$$

Let us now assume that the maximum terrain inclination over the whole earth satisfies

$$\beta_{\max} < 45^\circ ; \quad (47-29)$$

this will be the case after some suitable smoothing. Then, everywhere,

$$\tan \beta < 1 , \quad \eta = \frac{h-h_p}{l_0} < 1 ; \quad (47-30)$$

cf. Fig. 43.2. Hence the integrands in (47-28) will tend to zero as $r \rightarrow \infty$, and it may be shown that this implies that the integrals also tend to zero

(this is nontrivial since the integrals are singular), more precisely that

$$\lim_{r \rightarrow \infty} \|B_{2r}\| = 0. \quad (47-31)$$

From this it directly follows that the norms of all operators B_{2r} are uniformly bounded by a number M , so that

$$\|B_{2r}\| \leq M < \infty. \quad (47-32)$$

The absolute convergence of a series of linear operators is equivalent to the convergence of the series of their norms; hence (47-26) will converge if

$$\sum_{n=0}^{\infty} k^{2r} \|B_{2r}\| \quad (47-33)$$

converges, which is the case for $k < 1$. In fact, a majorant of this series is

$$\sum_{r=0}^{\infty} k^{2r} M = \frac{M}{1-k^2}, \quad (47-34)$$

which converges for $k < 1$.

Similar expansions hold for the operator ϕ_k^n , which is the n -th power of ϕ_k . We put

$$\phi_k^n = \sum_{r=0}^{\infty} k^{2r} B_{2r}^{(n)}. \quad (47-35)$$

Now a majorant of this series is the n -th power of the series (47-34), namely

$$M^n \sum_{r=0}^{\infty} (-1)^r \binom{-n}{r} k^{2r} = \frac{M^n}{(1-k^2)^n}, \quad (47-36)$$

which likewise converges for $k < 1$. Therefore, all the series (47-35) will converge for $k < 1$.

By multiplying all elevations by a factor q very slightly greater than 1 (so that the new θ_{\max} is still $< 45^\circ$) and by dividing k by the same q , it may be easily shown that these series will converge even for $k = 1$; cf. (Moritz, 1972, pp.12-13).

If we substitute all these series (47-35) into Neumann's series (47-18) and if we order with respect to equal powers of k , then we obtain

$$(I - k\Phi_k)^{-1} = I + k\psi_1 + k^2\psi_2 + \dots = \sum_{n=0}^{\infty} k^n \psi_n, \quad (47-37)$$

in which the operators ψ_n are independent of k .

By (47-17) we have

$$\kappa = \sum_{n=0}^{\infty} k^n \psi_n \Delta g, \quad (47-38)$$

which is (47-22) with

$$\kappa_n = \psi_n \Delta g. \quad (47-39)$$

If the operator series (47-37) converges, then the series (47-38), which is identical to (47-22), will converge as well.

This convergence of (47-37) and hence of Molodensky's series (47-22) follows by an application of Weierstrass' theorem on double series (Knopp, 1964, pp.444-445). This theorem says: Let a series

$$\sum_{n=0}^{\infty} f_n(k) \quad (47-40)$$

converge uniformly in k for $k \leq \rho$, for any $\rho < \rho_1$. Then the substitution

$$f_n(k) = \sum_{p=0}^{\infty} a_{np} k^p \quad (47-41)$$

into (47-40) and the subsequent rearrangement with respect to k leads to a power series which converges for $k < \rho_1$ provided the series (47-41) also converge for $k < \rho_1$.

Apply now this theorem with

$$f_n(k) = k^n \phi_k^n, \quad a_{np} = B_{2r}^{(n)}, \quad p = 2r; \quad (47-42)$$

the resulting series is then (47-37). We may work with power series whose "coefficients" are operators, belonging to some Banach space of linear operators, in exactly the same way as with power series whose coefficients are real or complex numbers; cf. (Dieudonné, 1960, chapter IX). The result is the following basic

THEOREM. Let the positive number C be defined by (47-20) and let $\beta_{\max} < 45^\circ$. Then Molodensky's series (47-22) converges for $k < k_0 = 1/C$ if $C \geq 1$, and it converges uniformly for $k \leq 1$ if $C < 1$.

Hence, even if $\beta_{\max} < 45^\circ$, the convergence of Molodensky's series at the earth's surface (for $k = 1$) is assured only if $C < 1$, otherwise convergence will hold only for $k < k_0$.

As we shall see later, we shall unfortunately not be able to guarantee $C < 1$. Therefore the essence of this theorem is that the Molodensky series (47-22) converges for the same values of k as the Neumann series (47-18), provided that $\beta_{\max} < 45^\circ$.

Molodensky's series for T . Let us now link the present results with Brovar's method described in sec. 44. By (47-2) we have

$$\mu = \kappa \cos^{-2} \beta = \kappa (1 + k^2 \tan^2 \beta). \quad (47-43)$$

Substituting the series (47-22) and rearranging with respect to equal powers of k gives a series

$$\mu = \sum_{p=0}^{\infty} k^p \mu_p \quad (47-44)$$

which is nothing else than (44-30). This series clearly converges whenever (47-22) does.

This shows the convergence of Molodensky's series for μ as given by (47-44) (it is a Molodensky series in the sense of having been obtained by Molodensky's method of expanding with respect to the parameter k).

To get the Molodensky series for T according to the procedure of sec. 44 (pp. 373-375), we substitute the series (47-44) into the Brovar representation (44-41) for T ,

$$T = \frac{1}{4\pi} \iint_{\sigma} \mu \left[S(r_p, \psi, r) - \frac{1}{r_p} \right] R^2 d\sigma,$$

and expand Stokes' function by (44-50) and (44-51). The formal multiplication of the power series with respect to k then gives (44-53), which is Molodensky's series for T :

$$T = \sum_{n=0}^{\infty} k^n T_n. \quad (47-45)$$

It is highly probable that the series (47-45) converges together with (47-44), although a rigorous proof has not been given so far.

In this way we see that the convergence of Molodensky's series for μ , and most likely also for T , is indeed determined by the convergence of the Neumann series (47-18). The condition $\beta_{\max} < 45^\circ$ unfortunately is by no means sufficient for the convergence of Molodensky's series, contrary to what is sometimes found in the literature.

The L_2 norm. Thus the problem reduces to estimating the norms $\|\phi_k\|$ or the number C in (47-20). For computing actual estimates, the L_2 norm, defined by

$$\|f\|^2 = \iint_{\sigma} f^2 d\sigma, \quad (47-46)$$

is most convenient. We take $R = 1$, which only corresponds to a particular choice of the unit of length; furthermore we first put $k = 1$. Then (47-13) becomes

$$\phi f = \frac{1}{2\pi} \iint_{\sigma} l^{-3} l_{0n} (1 + \tan^2 \beta) f d\sigma, \quad (47-47)$$

which we split up in the form

$$\phi f = \phi_1 f + \phi_2 f, \quad (47-48)$$

$$\phi_1 f = \frac{1}{2\pi} \iint_{\sigma} l^{-2} A f^* d\sigma, \quad (47-49)$$

$$\phi_2 f = \frac{1}{2\pi} \iint_{\sigma} (l^{-3} l_{0n} - l^{-2} A) f^* d\sigma, \quad (47-50)$$

$$f^* = f(1 + \tan^2 \beta), \quad (47-51)$$

$$A = \frac{\tan \beta_p \cos \alpha}{(1 + \tan^2 \beta_p \cos^2 \alpha)^{3/2}}. \quad (47-52)$$

Here l_0 is given by (47-16), l by (47-15) with $k = 1$, and n by (47-30); β_P is the maximum terrain inclination at the computation point P , and α is the azimuth from P to $d\sigma$, counted from the direction of maximum inclination.

The first integral $\phi_1 f$ has a standard form of a singular integral and can be estimated by a formula of Calderon-Zygmund type. The result is

$$\|\phi_1\| \leq \frac{1}{\pi} \tan \beta_{\max} (1 + \tan^2 \beta_{\max}) . \quad (47-53)$$

Geometrically, ϕ_1 is obtained from ϕ by replacing the terrain by its tangential plane at P (this is strictly true only if the reference surface is a plane instead of a sphere). The term $\phi_1 f$ has, therefore, the same strong singularity as ϕf , so that in the remaining part $\phi_2 f = \phi f - \phi_1 f$ this strong singularity cancels and only a weak singularity remains.

Although, for $\beta_{\max} = 45^\circ$, eq. (47-53) gives

$$\|\phi_1\| \leq 0.64 < 1 , \quad (47-54)$$

the corresponding estimate for $\|\phi\|$ comes out much larger. Detailed computations can be found in (Moritz, 1972, 1973b). The following estimate has been derived there:

$$\|\phi\| \leq \|\phi_1\| + \|\phi_2\| \leq \|\phi_1\| + \frac{1+9\tan^2 \beta_{\max}}{\cos^5 \beta_{\max}} \frac{R}{\rho_{\min}} , \quad (47-55)$$

where ρ_{\min} is the smallest radius of curvature of any normal profile of the terrain. Obviously ρ_{\min} will be much smaller than the earth's radius R , depending on the degree of smoothing applied to the topographic surface.

The estimate (47-55) also holds uniformly for all ϕ_k , $0 \leq k \leq 1$. It is thus an estimate for the constant C defined by (47-20) and occurring in the theorem given above. In this way we can assert convergence of Molodensky's series only for small values $k < k_0$ but not for $k = 1$, for the actual earth's surface: for this we should need $\|\phi\| < 1$.

It should, however, be noted that the estimate (47-55) is probably far too pessimistic, in the sense that a realistic estimate for $\|\phi\|$ will be much lower. This is a general feature of mathematical estimates, which always estimate "from above" and are thus heavily biased in a "pessimistic" sense.

It is, therefore, important to interpret our result correctly. We have not proved that Molodensky's series diverges at the earth's surface. Perhaps it converges for $k = 1$, but we have not been able to show this: we have proved convergence only for very small k .

It is clearly seen that the estimate for the operator norm $\|\Phi\|$ given by the expression (47-55) depends in an essential way on the maximum terrain inclination β_{\max} and on the maximum curvature $1/\rho_{\min}$. Thus the smoothness of topography plays a basic role in questions of convergence. We also see that a well-defined smoothing of the topography, which removes very steep slopes and sharp ridges, is essential before the convergence question can even be put in a meaningful way; it can be solved only for very smooth topography (corresponding to small values of k).

Convergence for $k < k_0$ may seem commonplace, but it does not hold, for instance, in certain series of celestial mechanics to which we shall come back at the end of this section. Thus, the solution given by Molodensky's series is stable at least for a sufficiently smooth topography, which is by no means trivial.

Hölder norms. Convergence in the L_2 norm does not necessarily imply pointwise convergence. To get pointwise and even uniform convergence for all points of the sphere, it would be appropriate to use Hölder norms.

A continuous function $f(P)$ on the sphere σ satisfies a Hölder condition, or Lipschitz condition, with exponent α if the expression

$$\sup_{P, Q \in \sigma} \frac{|f(P) - f(Q)|}{|PQ|^\alpha} \quad (47-56)$$

is finite, "sup" denoting the supremum or least upper bound. Here α is a positive number ≤ 1 and $|PQ|$ denotes the spherical distance between the points P and Q situated on the sphere. Such functions form a Banach space H^α ; the norm is given by

$$\|f\|_\alpha = \max_{P \in \sigma} |f(P)| + \sup_{P, Q \in \sigma} \frac{|f(P) - f(Q)|}{|PQ|^\alpha}. \quad (47-57)$$

Let f_n , $n = 1, 2, 3, \dots$, be a sequence approximating f such that

$$\lim_{n \rightarrow \infty} \|f - f_n\|_\alpha \rightarrow 0. \quad (47-58)$$

From it then follows that also

$$|f - f_n| \rightarrow 0$$

(47-59)

uniformly on the sphere, so that convergence in H^a implies pointwise, and even uniform, convergence.

It can be shown that singular integrals of type (47-49) are continuous not only in L_2 , but also in Hölder space H^a ; the same holds for (47-44) (Mikhlin, 1965, §6; Miranda, 1970, chapter II). Thus we can say that *Molodensky's series converges, for sufficiently small k , also in H^a and hence uniformly on the sphere*. However, a reasonable numerical estimate for the constant C in our convergence theorem seems to be even more difficult to get than in the L_2 case.

Practical implications of the convergence problem. Convergence has a different meaning for the mathematician and for the geodesist who does numerical computations. For the mathematician, convergence holds if a condition such as (47-59) is satisfied, even if the convergence is very "slow", that is, if a practically usable approximation is only achieved by taking n very high. For the numerical geodesist, however, convergence means *rapid convergence*, in which f_n furnishes a good approximation to f already for low n , so that the first few terms of a series already give a practically sufficient accuracy.

It thus appears as if "mathematical convergence" would be a necessary (though not sufficient) condition for "numerical convergence". Curiously enough, this is not the case, as Poincaré (1893, chapter VIII) has pointed out in a detailed fashion.

The following example, taken from (Erdélyi, 1956), has already been discussed by Euler in 1754. The series

$$S(x) = 1 - 1!x + 2!x^2 - 3!x^3 + - \dots \quad (47-60)$$

is certainly divergent for all $x \neq 0$. However, for small x (say 10^{-2}) the terms of the series at first decrease rapidly, and we can try to compute an approximate value of the function $S(x)$ by using the first few terms only. Later on, the series terms will increase, but if we neglect the higher terms and use only the first few ones, we still get a quite accurate estimate of the values of a certain analytical function which is the formal "sum" of (47-60).

Expansions of this kind have been called *semi-convergent* or *convergently beginning series*; today the name *asymptotic series* is used. Thus divergent series can be "numerically convergent"!

According to Poincaré, many series of celestial mechanics are asymptotic series. They are perfectly suited for numerical computations, although mathematically divergent.

Also Molodensky's series can be treated as an asymptotic expansion with respect to k (Moritz, 1971). As the argument of the present section shows, we are here slightly better off: our series converges at least uniformly for very small values of k ; not even this holds for the series of celestial mechanics.

We now recognize that the question whether Molodensky's series converges or diverges for the actual earth's surface ($k = 1$) is largely irrelevant for practical applications. If it did converge, but very slowly, Molodensky's series would be practically useless. It is easy to see that data errors (in Δg and h) become increasingly effective in the higher-order terms; it appears hardly meaningful to go much beyond $n = 2$. On the other hand, test computations show that taking $n = 2$ or even only $n = 1$, we get practically quite satisfactory results. Thus Molodensky's series is "convergently beginning", and this is what we need in practice.

48. USE OF THE TERRAIN CORRECTION

In sec. 45 we have seen that we may obtain the height anomaly ζ and the components ξ, η of the deflection of the vertical by the formulas (45-42) and (45-43). If we retain only the first-order correction g_1 , neglecting terms of higher order, we have the gradient solution (45-50):

$$\zeta = \frac{R}{4\pi\gamma} \iint_{\sigma} (\Delta g + g_1) S(\psi) d\sigma, \quad (48-1)$$

$$\begin{Bmatrix} \xi \\ \eta \end{Bmatrix} = \frac{1}{4\pi\gamma} \iint_{\sigma} (\Delta g + g_1) \frac{dS}{d\psi} \begin{Bmatrix} \cos\alpha \\ \sin\alpha \end{Bmatrix} d\sigma; \quad (48-2)$$

we have written γ instead of γ^0 . Briefly this is

$$\zeta = S(\Delta g + g_1), \quad (48-3)$$

$$\underline{\xi} = \underline{V}(\Delta g + g_1), \quad (48-4)$$

S denoting Stokes' operator and \underline{V} denoting Vening Meinesz' operator, $\underline{\xi}$ being an abbreviation of the pair $[\xi, \eta]$.

In the present section it will be shown that the formulas

$$\zeta = S(\Delta g + C) , \quad (48-5)$$

$$\xi = \underline{V}(\Delta g + C) , \quad (48-6)$$

are as accurate as (48-3) and (48-4), *provided the gravity anomaly is linearly dependent on the elevation* h . Here C denotes the well-known *terrain correction* which is easier to compute than g_1 ; cf. (Heiskanen and Moritz, 1967, p.131). This idea is due to Pellinen (1962); for a physical interpretation cf. (Moritz, 1968b).

The correction g_1 is given by (45-51) and (45-48) as

$$(g_1)_P = - (h_P - h_A) \frac{R^2}{2\pi} \iint_{\sigma_Q} \frac{\Delta g_Q - \Delta g_P}{l_{PQ}^3} d\sigma_Q ; \quad (48-7)$$

here we have consistently used subscripts to designate:

- A ... point at which ξ , η , ζ are computed;
- P ... point to which Δg and g_1 in Stokes' and Vening Meinesz' formulas refer;
- Q ... point to which the surface element $d\sigma$ in (48-7) refers.

Obviously,

$$l_{PQ} = (l_O)_{PQ} = 2R \sin(\psi_{PQ}/2) , \quad (48-8)$$

ψ_{PQ} denoting the spherical distance between P and Q .

The terrain correction may be expressed by

$$C_P = \frac{1}{2} G \rho R^2 \iint_{\sigma_Q} \frac{(h_Q - h_P)^2}{l_{PQ}^3} d\sigma_Q , \quad (48-9)$$

G denoting the gravitational constant and ρ the density of the topographic masses, assumed constant. The derivation of this expression is left to the reader (hint: expand eq.(3-14) of (Heiskanen and Moritz, 1967) for small $b = |h - h_P|$ and let the size of compartments in (3-20), *ibid.*, become infinitesimal). Then (48-5) and (48-6) become

$$\zeta_A = S_{AP}(\Delta g + C)_P, \quad \xi_A = V_{AP}(\Delta g + C)_P. \quad (48-10)$$

To show the equivalence, we thus have to compute

$$S_{AP}(C - g_1)_P \quad \text{and} \quad V_{AP}(C - g_1)_P. \quad (48-11)$$

The gravity anomalies Δg are defined by (42-15); they are free-air anomalies referred to the earth's surface; cf. (Heiskanen and Moritz, 1967, p.293). Let us assume that they are linearly correlated with elevation, so that there holds a relation

$$\Delta g = a + bh, \quad (48-12)$$

where a and b are approximately constants (*ibid.*, pp.283-284). The constant b is given by

$$b = 2\pi G\rho, \quad (48-13)$$

whereas a is the slowly varying Bouguer anomaly which may be considered locally constant. Let us, however, assume now that (48-12) holds exactly with constant a and b , so that Δg is linearly dependent on h .

Then (48-7) becomes

$$(g_1)_P = -2\pi G\rho \frac{R^2}{2\pi} \iint_{\sigma} \frac{(h_P - h_A)(h_Q - h_P)}{l_{PQ}^3} d\sigma_Q. \quad (48-14)$$

This is subtracted from (48-9) with the result

$$(C - g_1)_P = \frac{1}{2} G\rho R^2 \iint_{\sigma} \left[(h_Q - h_P)^2 + 2(h_P - h_A)(h_Q - h_P) \right] l_{PQ}^{-3} d\sigma_Q, \quad (48-15)$$

which is easily seen to be equal to

$$(C - g_1)_P = \pi G\rho \frac{R^2}{2\pi} \iint_{\sigma_Q} \frac{h_Q^2 - h_P^2}{l_{PQ}^3} d\sigma_Q - 2\pi G\rho h_A \frac{R^2}{2\pi} \iint_{\sigma_Q} \frac{h_Q - h_P}{l_{PQ}^3} d\sigma_Q \quad (48-16)$$

or

$$C - g_1 = \pi G \rho L(h^2) - 2\pi G \rho h_A L(h) , \quad (48-17)$$

using the operator L defined by (45-31). Thus

$$S(C - g_1) = \pi G \rho S L(h^2) - 2\pi G \rho h_A S L(h) , \quad (48-18)$$

$$\underline{V}(C - g_1) = \pi G \rho \underline{V} L(h^2) - 2\pi G \rho h_A \underline{V} L(h) . \quad (48-19)$$

Now by (46-30) there is

$$S L = -\gamma^{-1} I ; \quad (48-20)$$

we have the factor γ^{-1} since Stokes' operator now gives $\zeta = \gamma^{-1} T$ instead of T ; I denotes the unit operator. To find $\underline{V} L$, we keep in mind that for Stokes' problem

$$\underline{\xi} = \underline{V}(\Delta g) = -\text{grad} \zeta = -\text{grad} S(\Delta g) , \quad (48-21)$$

where the symbol grad denotes the operator $[\partial/\partial x, \partial/\partial y]$. Hence

$$\underline{V} = -\text{grad} \cdot S , \quad (48-22)$$

whence, by (48-20),

$$\underline{V} L = -\text{grad} \cdot S \cdot L = \gamma^{-1} \text{grad} . \quad (48-23)$$

Using (48-23) we thus find

$$\gamma \underline{V}(C - g_1) = \pi G \rho \text{grad}(h^2) - 2\pi G \rho h_A \text{grad} h . \quad (48-24)$$

Now

$$\begin{aligned} \text{grad} h^2 &= \left(\frac{\partial h^2}{\partial x} , \frac{\partial h^2}{\partial y} \right) \\ &= 2h \left(\frac{\partial h}{\partial x} , \frac{\partial h}{\partial y} \right) = 2h \text{grad} h . \end{aligned} \quad (48-25)$$

In (48-24), all quantities refer to A , so that

$$\gamma \underline{V}(C - g_1) = \pi G \rho 2h_A (\text{grad } h)_A - 2\pi G \rho h_A (\text{grad } h)_A = 0 \quad (48-26)$$

whence

$$\underline{V}(C) = \underline{V}(g_1) . \quad (48-27)$$

This proves the exact equivalence of the Vening Meinesz formulas (48-4) and (48-6) under the conditions mentioned.

For the corresponding Stokes' formulas we have (48-18) which by (48-20) becomes

$$\gamma S(C - g_1) = -\pi G \rho h_A^2 + 2\pi G \rho h_A \cdot h_A = \pi G \rho h_A^2 , \quad (48-28)$$

which is not exactly zero. Thus

$$S(C) = S(g_1) + \pi G \rho \gamma^{-1} h_A^2 . \quad (48-29)$$

To get a numerical estimate of the last term in (48-29), we take $\rho = 2.67 \text{ g/cm}^3$ and $h_A = 1000 \text{ m}$; then

$$\pi G \rho \gamma^{-1} h_A^2 \doteq 5 \text{ cm} . \quad (48-30)$$

It will be consistent with the present approximation--restriction to first-degree corrections to Stokes' and Vening Meinesz' formulas and assumption of linear dependence of Δg on h --to neglect this small term in (48-29), so that

$$S(C) \doteq S(g_1) . \quad (48-31)$$

Thus (48-1) and (48-2) become in view of (48-27) and (48-31):

$$\zeta = \frac{R}{4\pi\gamma} \iint_{\sigma} (\Delta g + C) S(\psi) d\sigma , \quad (48-32)$$

$$\begin{Bmatrix} \xi \\ \eta \end{Bmatrix} = \frac{1}{4\pi\gamma} \iint_{\sigma} (\Delta g + C) \frac{dS}{d\psi} \begin{Bmatrix} \cos\alpha \\ \sin\alpha \end{Bmatrix} d\sigma , \quad (48-33)$$

which proves (48-5) and (48-6). To repeat, (48-33) follows rigorously from (48-2) and (48-12), whereas (48-32) involves the neglect of the small term (48-28).

The sum $\Delta g + C$, free-air anomaly plus terrain correction, is sometimes called *Faye anomaly*. Thus we may say that the use of the Faye anomaly in Stokes' and Vening Meinesz' formula gives a better approximation for Molodensky's problem than the use of the uncorrected free-air anomaly Δg .

49. PRACTICAL ASPECTS

Formulas to be used. The practical solution of Molodensky's problem consists in adding certain correction terms to the formulas of Stokes and Vening Meinesz. These Molodensky corrections are of particular importance in calculating deflections of the vertical ξ, η .

In fact, loosely speaking, Stokes' formula is about ten times as accurate as Vening Meinesz' formula, in the following sense. An error of 1" in ξ corresponds to an error of 30 m in position. An accuracy of $\pm 0.3'' \approx \pm 10$ m in ξ or η may be about as difficult to get as an accuracy of ± 1 m in ζ , which means that the gravimetric method gives vertical position by about one order of magnitude more accurately than horizontal position.

A similar, but even more pronounced, relation holds between the orders of magnitude of the Molodensky corrections for ζ on the one hand and ξ and η on the other hand. In a topography that ranges from flat to mountainous such as West Germany, the Molodensky correction in ζ has generally the order of a few centimeters, whereas the correction of ξ and η may reach at least a few tenths of a second of arc, that is, about 10 m; cf. (Groten, 1979, vol.2, pp.685-686).

Thus, the computational formulas should be selected with particular regard to the deflection of the vertical. From this respect, the solution by analytical continuation (sec.45) is preferable to Molodensky's (sec.43) or Brovar's (sec.44) solution. In fact, Molodensky's solution (the same holds for Brovar's solution) for vertical deflections is more complicated and contains two relatively large terms of the same order of magnitude but of opposite sign, which are computed in quite different ways, whereby additional errors may be introduced; the gradient solution is free from this deficiency (Heiskanen and Moritz, 1967, sec.8-9).

To second order, the solution by analytical continuation as given by (45-42), (45-43), and (45-49) becomes

$$\zeta = \frac{R}{4\pi\gamma_0} \iint_{\sigma} \Delta g' S(\psi) d\sigma, \quad (49-1)$$

$$\begin{Bmatrix} \xi \\ \eta \end{Bmatrix} = \frac{1}{4\pi\gamma_0} \iint_{\sigma} \Delta g' \frac{dS}{d\psi} \begin{Bmatrix} \cos\alpha \\ \sin\alpha \end{Bmatrix} d\sigma, \quad (49-2)$$

$$\Delta g' = \Delta g + g_1 + g_2,$$

$$g_1 = -(h - h_A) L_1(\Delta g), \quad (49-3)$$

$$g_2 = -(h - h_A)^2 L_2(\Delta g) - (h - h_A) L_1(g_1) \quad (49-4)$$

and

$$L_1(f) = \frac{R^2}{2\pi} \iint_{\sigma} \frac{f - f_P}{l_0^3} d\sigma. \quad (49-5)$$

The term $L_2(\Delta g)$ may be evaluated by an iteration of L_1 :

$$L_2(\Delta g) = \frac{1}{2} L_1[L_1(\Delta g)] \quad (49-6)$$

or directly by (45-34):

$$L_2(\Delta g) = -\frac{1}{2} \left[\frac{\partial^2 \Delta g}{\partial x^2} + \frac{\partial^2 \Delta g}{\partial y^2} \right]. \quad (49-7)$$

If only ζ is computed, then g_2 can certainly be neglected. If ξ , η , ζ are computed in mountains, then g_2 may have an effect on the deflection of the vertical and may be added, together with g_1 , to Δg to give the gravity anomaly to be used in Stokes' and Vening Meinesz' formulas (49-1) and (49-2).

It may not be feasible to compute corrections of higher than the second order even if they would have an effect in extreme situations, because the influence of errors in the data Δg on these higher terms might increase so strongly as to render their numerical values almost meaningless.

Much simpler and practically sufficient in many cases is the solution obtained by neglecting g_2 , so that

$$\Delta g' = g + g_1 = \Delta g - (h - h_A) L_1(\Delta g). \quad (49-8)$$

This is the *gradient solution* (45-50) described in (Heiskanen and Moritz, 1967, eqs.(8-71) and (8-80)).

Still easier for computational purposes is the *terrain correction solution*; it consists in using the Faye anomaly

$$\Delta g_F = \Delta g + C \quad (49-9)$$

instead of $\Delta g'$ in the simple formulas of Stokes and Vening Meinesz, C being the usual terrain correction. As we have seen in the preceding section, this *terrain correction solution* is equivalent to the gradient solution provided the gravity anomaly is linearly dependent on the elevation. Since there will be, at best, only a statistical correlation of Δg with h , rather than an exact functional relation, this equivalence will only be approximate. We may say that the terrain correction solution will have an intermediate accuracy between the gradient solution, using (49-8), and the crude Stokes' and Vening Meinesz' formulas, neglecting all corrections.

Other linear solutions (Moritz, 1968a) seem less practical.

Computational considerations. In the gradient solution given by equations (49-1) through (49-6), all computations are reduced to the evaluation of three integrals: Stokes' and Vening Meinesz' integrals and the gradient integral (49-5). The practical evaluation of these three integrals has been discussed, e.g., in (Heiskanen and Moritz, 1967, sec.2-24); a mathematical investigation has been made by Meissl (1971c).

For automatic computation, one considers blocks of various sizes bounded by geographical grid lines: $5^\circ \times 5^\circ$, $1^\circ \times 1^\circ$, $30' \times 30'$, down to $5' \times 5'$ and smaller; mean values of Δg in such blocks are stored. To get approximately square blocks, one might also use $20' \times 30'$ and similar sizes.

In the neighborhood of the computation point it is appropriate to use smaller blocks than for distant zones. Recent suggestions have been given by Lelgemann (1974) and Clarke (1978). For Stokes' integral, Lelgemann uses $6' \times 10'$ blocks up to a distance $\psi \leq 1^\circ$ from the computation point, then $12' \times 20'$ and $20' \times 30'$ up to $\psi = 5^\circ$, then $1^\circ \times 1^\circ$ up to $\psi = 20^\circ$, and global spherical-harmonic representation of Δg outside 20° . Clarke has a similar pattern: $6' \times 6'$ up to $\psi = 1.5^\circ$, then $30' \times 30'$ to $\psi = 5^\circ$, then $1^\circ \times 1^\circ$ to $\psi = 20^\circ$, and $5^\circ \times 5^\circ$ for $\psi > 20^\circ$. Such a pattern can also be used for Vening Meinesz' integral, depending on the accuracy to be attained and on the gravity data available.

Essential is the effect of the innermost zone, especially in Vening Meinesz' formula and in the gradient integral, which have a stronger singularity than Stokes' integral and, therefore, require more detail around the

computation point. In this zone, say of size $30' \times 30'$, one might use a detailed representation of Δg by a bicubic spline approximation (Sünkel, 1977); see also (Moritz, 1978b). Since spline functions are piecewise polynomials, the necessary integration in the three basic integral formulas can be performed analytically.

On the other hand, the effect of the zone beyond about 50 km from the computation point seems to be negligible in the gradient integral.

Of particular importance, especially if one employs the analytical continuation solution to higher orders, is the use of a consistent field of gravity anomalies Δg and elevations h , both, e.g., in the form of a spline representation. This avoids the use of values from different representations which might be incompatible. It is clear that such gravity and terrain models involve some smoothing which is appropriate in Molodensky's problem (also in order to obtain practical convergence!) but must be done in a consistent fashion.

For similar reasons, the use of actual observations of the vertical gradient $\partial \Delta g / \partial h$ appears less advisable than the computation of $L_1(\Delta g)$ from the given Δg -field: actual observations tend to be too irregular and may not be compatible with the Δg -field.

For the inner zone, least-squares interpolation may also be used (Lachapelle, 1977), thus getting an approach that combines integral formulas and collocation. For other aspects of such a combination cf. (Moritz, 1976a).

Effect of the atmosphere. The theory of Molodensky presupposes that the anomalous potential T outside the earth's surface is a harmonic function or, in other terms, that the space outside the earth is empty. Thus the effect of the atmosphere on Δg must be removed by computation.

In view of the extreme smallness of atmospheric effects, a very simple model is sufficient for this purpose. We regard the reference ellipsoid formally as a sphere (spherical approximation) and assume that the atmosphere consists of spherically symmetric layers (Fig.49.1). For the sake of simplicity it is preferable to remove the atmosphere not only above the earth's surface, but formally everywhere above the reference sphere.

The vertical attraction F of the atmosphere at a point P is the sum of the attraction F_1 of the atmospheric layer A_1 between the spheres E and K and the attraction F_2 of the atmospheric layer A_2 outside the sphere K (Fig.49.1). It is well known from potential theory that the attraction of a spherical shell on a point inside the shell is zero. From this follows in view of the spherical symmetry of the atmosphere that $F_2 = 0$. For F_1 , however, there follows, again from the spherical symmetry, that the attraction is rigorously given by the Newtonian law for a

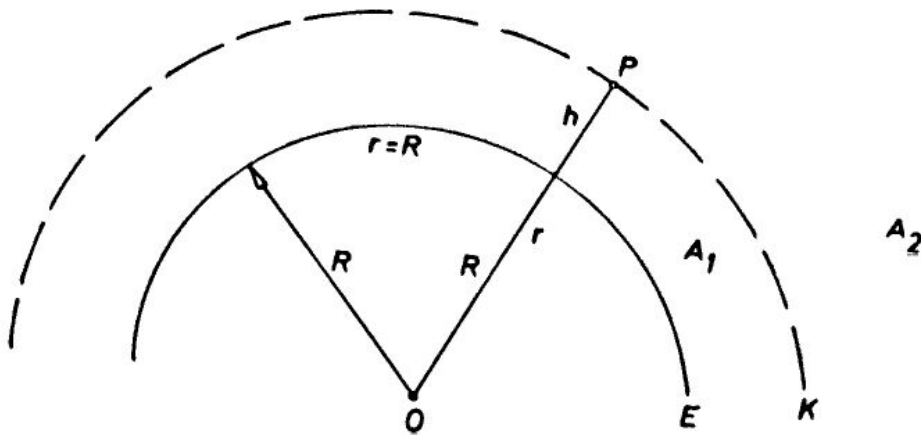


FIGURE 49.1. A spherically symmetric atmosphere.

point mass, since the external gravitational field of a spherically symmetric body is the same as if the mass of the body were concentrated at its center. Thus

$$F = F_1 = G \frac{m(r)}{r^2}, \quad (49-10)$$

where $m(r)$ is the mass of the atmospheric layer A_1 as a function of the radius r .

It is now convenient to introduce the mass $M(r)$ of the atmosphere A_2 outside the sphere K through P ; there is obviously

$$m(r) = M_A - M(r), \quad (49-11)$$

M_A being the total mass of the atmosphere. Then we get from (49-10):

$$F = G \frac{M_A}{r^2} - G \frac{M(r)}{r^2}. \quad (49-12)$$

The first term on the right-hand side can be regarded as the attraction of the total mass of the atmosphere, transported into the earth's interior and distributed there in a spherically symmetric way. This term is automatically taken into account if the mass of the reference ellipsoid includes the mass of the atmosphere, which is the usual (and the most appropriate) treatment; it is followed, e.g., in the Geodetic Reference System 1967 (IAG, 1970).

Thus there remains only the second term, and this is the *atmospheric effect on gravity*:

$$\delta g_A = G \frac{M(r)}{r^2} ; \quad (49-13)$$

the plus sign corresponds to the removal of the atmospheric effect.

If $\rho(r)$ represents the spherically symmetric atmospheric density law, then the integration over A_2 , dv being the volume element, gives

$$M(r) = \iiint_{A_2} \rho(r) dr = 4\pi \int_r^\infty \rho(r) r^2 dr . \quad (49-14)$$

The integration with respect to r can be performed if the law $\rho(r)$ is taken from a standard atmosphere. In this way δg_A has been computed and tabulated in (IAG, 1970); there an ellipsoidal density distribution has been assumed, but a spherical one is practically as accurate.

The corresponding effect of the atmosphere on the gravity potential W follows from an integration of (49-13):

$$\delta W_A = \int_\infty^r \delta g_A dr = -G \int_r^\infty \frac{M(r)}{r^2} dr . \quad (49-15)$$

This potential change causes a shift of the level surfaces by

$$\delta \zeta_A = \frac{\delta W_A}{\gamma} , \quad (49-16)$$

which is a consequence of Bruns' formula (42-16).

Thus the atmospheric effect is taken into account by the following procedure:

1. Add to the measured gravity the positive correction (49-13), in which

$$r = R + h , \quad (49-17)$$

h being the height of the observation point.

2. Form the gravity anomaly

$$\Delta g = g + \delta g_A - \gamma , \quad (49-18)$$

where γ refers to the telluroid, and apply Molodensky's theory; denote the resulting height anomaly by ζ' .

3. The real height anomaly ζ is then

$$\zeta = \zeta' + \delta\zeta_A \quad (49-19)$$

where $\delta\zeta_A$ is given by (49-16).

The corrections δg_A , δW_A , and $\delta\zeta_A$ may be tabulated as a function of the height h ; thus the atmospheric corrections are very easy to apply. To give an idea of the order of magnitude, we mention that at sea level $\delta g_A = 0.87 \text{ mgal}$ and $\delta\zeta_A = -0.7 \text{ cm}$.

In view of the accuracy feasible at present and in the near future, $\delta\zeta_A$ can safely be neglected. Thus the atmospheric correction reduces to taking δg_A , as tabulated on pp. 72-73 of (IAG, 1970) as a function of the observation station, and adding it to the measured gravity g .

Ellipsoidal corrections. Since the reference ellipsoid is very nearly a sphere, an expansion in terms of e^2 , the square of the excentricity, is appropriate. Such expansions (restricted to terms linear in e^2) were employed in sec. 39. We shall use these developments to derive ellipsoidal corrections to Stokes' and Vening Meinesz' formulas.

By (39-6) and (39-78) we have

$$T = T^O, \quad (49-20)$$

$$\Delta g' = \Delta g^O + e^2 \Delta g^1, \quad (49-21)$$

The "spherical parts" are related by Stokes' formula

$$T^O = \frac{R_A}{4\pi} \iint_{\sigma} \Delta g^O S(\psi) d\sigma; \quad (49-22)$$

in fact, we have reduced Molodensky's problem to Stokes' problem for "point level" by replacing Δg by $\Delta g'$; cf. p. 379. The radius R_A is given by

$$R_A = R + h_A, \quad (49-23)$$

h_A being the height of the computation point A ; in keeping with the present higher accuracy requirements it may no longer be sufficient simply to replace R_A by R as we have done so far. The ellipsoidal correction Δg^1 is given by (39-80):

$$\Delta g^1 = \frac{1}{R} \sum_{n=2}^{\infty} \sum_{m=0}^n [G_{nm} R_{nm}(\theta, \lambda) + H_{nm} S_{nm}(\theta, \lambda)] , \quad (49-24)$$

the coefficients G_{nm} and H_{nm} being defined by (39-81) and (39-82). The substitution of (49-20) and (49-21) into (49-22) yields immediately

$$T = \frac{R_A}{4\pi} \iint_{\sigma} (\Delta g^1 - e^2 \Delta g^1) S(\psi) d\sigma . \quad (49-25)$$

Then the height anomaly ζ is given by (39-19), (39-20), and (39-21), noting that N is now ζ :

$$\zeta = \frac{R_A}{4\pi\gamma_A} \iint_{\sigma} (\Delta g^1 - e^2 \Delta g^1) S(\psi) d\sigma + e^2 \zeta^1 , \quad (49-26)$$

$$\zeta^1 = \left(\frac{1}{4} - \frac{3}{4} \sin^2 \phi \right) \zeta^0 , \quad (49-27)$$

ζ^0 being the first term on the right-hand side of (49-26). The symbol γ_A designates spherical gravity at elevation h_A :

$$\gamma_A = \gamma^0 + \frac{\partial \gamma^0}{\partial h} h_A , \quad (49-28)$$

$\partial \gamma^0 / \partial h$ denoting the normal spherical gravity gradient of -0.31 gal/km. The deflection components ξ, η are obtained analogously:

$$\begin{Bmatrix} \xi \\ \eta \end{Bmatrix} = \frac{1}{4\pi\gamma_A} \iint_{\sigma} (\Delta g^1 - e^2 \Delta g^1) \frac{dS}{d\psi} \begin{Bmatrix} \cos \alpha \\ \sin \alpha \end{Bmatrix} d\sigma + e^2 \begin{Bmatrix} \xi^1 \\ \eta^1 \end{Bmatrix} ; \quad (49-29)$$

ξ_1 and η_1 are expressed by (39-38) and (39-39), in which N^0 is ζ^0 and ξ^0 and η^0 are given by the first term on the right-hand side of (49-29).

It is sometimes more convenient to directly evaluate Stokes' integral over Δg^1 in (49-25) and (49-26); we simply have

$$\frac{R_A}{4\pi} \iint_{\sigma} \Delta g^1 S(\psi) d\sigma = \sum_{n=2}^{\infty} \frac{1}{n-1} \sum_{m=0}^n [G_{nm} R_{nm}(\theta, \lambda) + H_{nm} S_{nm}(\theta, \lambda)] . \quad (49-30)$$

This is a direct consequence of the well-known formula

$$T_n = \frac{R}{n-1} \Delta g_n ,$$

relating Laplace harmonics of T and Δg ; cf. (Heiskanen and Moritz, 1967, p.97). It is clearly permissible here to replace R_A by R .

A similar transformation is possible for Vening Meinesz' formulas, but this involves horizontal derivatives of spherical harmonics. Thus, if ξ and n are computed, alone or in combination with ζ , then the formulas (49-26) to (49-29) are simplest.

As we have remarked in sec. 39, it is sufficient to compute the ellipsoidal corrections with one of the available truncated spherical-harmonic representations of the external gravitational field.

To get the full benefit of accuracy improvement from ellipsoidal corrections, the mapping from the ellipsoid to the sphere must be well defined. This implies that the radius R of the sphere must be given by (39-5):

$$R = \sqrt[3]{a^2 b} = 6371.0 \text{ km} , \quad (49-31)$$

mean gravity γ^0 by (39-16):

$$\gamma^0 = \gamma_a \left(1 + \frac{1}{4} e^2 \right) = 979.7 \text{ gal} , \quad (49-32)$$

and ϕ, λ must be geographical coordinates on the ellipsoid. These values are to be used in (49-23) and (49-28).

These refinements are, of course, only necessary in the principal ("zero-order") terms; in the correction (first-order) terms we may safely replace R_A by R and γ_A by γ^0 .

For references on ellipsoidal corrections in the geodetic boundary-value problem, from Sagrebin (1956) to Lelgemann (1970), see p.316. The present formulas are essentially due to Lelgemann, although the derivation is different.

According to Lelgemann, ellipsoidal corrections reach 0.5 m in ζ and 0.05" in ξ and n ; he has also given global maps for them. With ellipsoidal corrections, we may compute ζ to a decimeter or even centimeter accuracy, provided the necessary coverage of the earth by gravity data is available. Unfortunately we are still far from such a desirable situation.

50. EXISTENCE AND UNIQUENESS FOR THE LINEARIZED MOLODENSKY PROBLEM

We shall now investigate the existence and uniqueness of the solution of the linear Molodensky problem. As an introduction we first examine Stokes' problem.

The problem of Stokes. Stokes' problem is the boundary-value problem in its simplest form: given the gravity anomaly on a sphere, to determine the anomalous potential T on and outside the sphere, assuming T to be harmonic outside this sphere. The corresponding boundary condition is (2.33); since the radial direction is normal to the bounding sphere, the oblique-derivative problem reduces in this case to a problem involving normal derivatives, which is much simpler.

The general solution is given by Stokes' integral formula

$$T(\theta, \lambda) = T_0 + \frac{R}{4\pi} \iint_{\sigma} \Delta g S(\psi) d\sigma + T_1(\theta, \lambda), \quad (50-1)$$

which expresses T on the given sphere in terms of Δg on this sphere. Here T_0 is a fixed constant related to the mass of the earth, and

$$T_1(\theta, \lambda) = A_1 \sin \theta \cos \lambda + A_2 \sin \theta \sin \lambda + A_3 \cos \theta \quad (50-2)$$

is a spherical surface harmonic of the first degree. Polar distance θ and longitude λ are spherical coordinates, and A_1, A_2, A_3 are arbitrary constants which have the following physical interpretation (Heiskanen and Moritz, 1967, p.99). Let ξ_1, ξ_2, ξ_3 denote the rectangular coordinates of the earth's center of gravity, the origin being the center of the ellipsoid. Then, approximately,

$$A_i = \gamma^0 \xi_i, \quad (50-3)$$

where γ^0 denotes a mean value of gravity over the earth. Therefore, non-zero A_i mean that the center of the reference ellipsoid does not coincide with the earth's center of mass.

A necessary and sufficient condition for Stokes' problem to be solvable for continuous boundary values is that the function Δg does not contain spherical harmonics of the first degree. In other terms, Δg must be orthogonal to any harmonic function of the first degree $Y_1(\theta, \lambda)$:

$$\iint_{\sigma} \Delta g(\theta, \lambda) Y_1(\theta, \lambda) d\sigma = 0 ; \quad (50-4)$$

cf. (Heiskanen and Moritz, 1967, p.97). Since $Y_1(\theta, \lambda)$ contains three constants, this equation comprises, in fact, three independent conditions.

The solution (50-1) contains three free constants A_1, A_2, A_3 . The solution can be made unique by putting all $A_1 = 0$, which means that the first-degree harmonic (50-2) vanishes.

The fact that Δg must satisfy three conditions and that the solution (50-1) contains three free constants expresses the so-called *Fredholm alternative*; see below.

It should also be pointed out that a solution (50-1) with $A_1 \neq 0 \neq A_2$ is physically impossible, although it is mathematically valid as a solution of the boundary-value problem defined by $\Delta T = 0$ outside the sphere and by the boundary condition (2-33) on the sphere.

In fact, for T to be harmonic and zero at infinity, the centrifugal potential contained in both W and U must be equal, so as to drop out in $T = W - U$. This requires that the axis of the reference ellipsoid coincides with the earth's axis of rotation. If this common axis is taken as x_3 axis, then the centrifugal potential is

$$\frac{1}{2} \omega^2 (x_1^2 + x_2^2) . \quad (50-5)$$

Indeed, if the two axes were only parallel and separated by the vector¹

$$[\delta x_1, \delta x_2, 0] ,$$

then T would contain a term

$$\omega^2 (x_1 \delta x_1 + x_2 \delta x_2) \quad (50-6)$$

due to the difference of the two centrifugal potentials; this term and therefore T , would not be zero at infinity. The same would hold if the two axes were not parallel.

¹We often use the vector notation $\underline{a} = [a_1, a_2, a_3]$ without distinguishing between row and column vectors, provided no matrix operations are involved.

So the two rotation axes must coincide. Since the earth's rotation axis passes through the center of mass for physical reasons, and since the axis of the ellipsoid contains the center of the ellipsoid for reasons of symmetry, both centers must lie on the common axis, which is taken as the x_3 coordinate axis. This implies that the two centers can differ only in the x_3 coordinate, so that ξ_1 and ξ_2 , and therefore A_1 and A_2 by (50-3), must be zero.

Thus, if a solution (50-1) is to be physically meaningful, only A_3 can differ from zero, so that the solution for a rotating earth has, in reality, only one degree of freedom. Since $A_1 = A_2 = 0$, it is quite natural to take also $A_3 = 0$, thus letting the center of the reference ellipsoid coincide with the earth's center of mass.

The simple Molodensky problem. This is the linear Molodensky problem for a spherical reference surface (sec.42). We shall prove existence and uniqueness of the solution for this problem by establishing a one-to-one correspondence between the Stokes problem and the simple Molodensky problem.

Let us consider the telluroid Σ , on which the boundary condition (42-14) is defined, together with a sphere S' concentric to the reference sphere and such that Σ is completely inside S' (Fig.50.1). This sphere S' might be called *Brillouin sphere*, after the French scientist who proposed gravity reduction to a level surface completely outside the earth.

The function

$$F = r\Delta g \quad (50-7)$$

is well known to be a *harmonic function* in space, r being the variable radius vector of the point under consideration (*ibid.*, p.90). As the boundary values of Δg , and hence of F , are given on the surface Σ , we can compute F , and hence Δg , at every point outside Σ by solving an *external Dirichlet problem*, which is uniquely solvable for continuous boundary data (cf. Kellogg, 1929, p.314). In particular, this gives Δg at every point P_r between the surfaces Σ and S' --to be denoted by $\Delta g(r)$ --and on the Brillouin sphere itself--to be denoted by $\Delta g'$.

From the values of Δg along a radius it is straightforward to compute radial differences of the potential T : by (Heiskanen and Moritz, 1967, p.92) we have

$$\frac{\partial}{\partial r}(r^2 T) = -r^2 \Delta g(r), \quad (50-8)$$

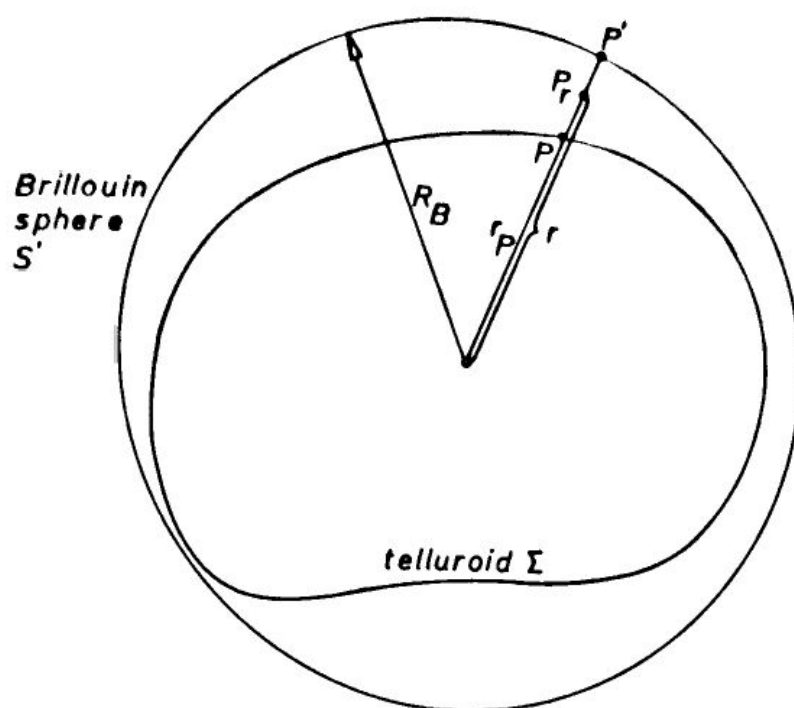


FIGURE 50.1. The Brillouin sphere.

which on integration gives

$$(r^2 T)_P - (r^2 T)_{P'} = \int_P^{P'} r^2 \Delta g(r) dr \quad (50-9)$$

or with the symbols of Fig. 50.1,

$$r_P^2 T - R_B^2 T' = \int_{r_P}^{R_B} r^2 \Delta g(r) dr, \quad (50-10)$$

T denoting the potential on the telluroid and T' on the Brillouin sphere. Now we can solve Molodensky's problem by the following three steps:

1. Computation of $\Delta g(r)$ and $\Delta g'$ by solving the external Dirichlet problem.
2. Determination of T' from $\Delta g'$ by solving Stokes' problem for the sphere S' .
3. Computation of T at Σ from T' at S' by (50-10).

Steps 1 and 3 are *one-to-one* because Dirichlet's problem is uniquely solvable and because different functions Δg on Σ correspond to different functions $\Delta g'$ on S' and vice versa. Thus the question of solvability

of Molodensky's problem for the telluroid Σ has been reduced to the question of solvability of Stokes' problem for the sphere S' , to which the answer has been given above. For the simple Molodensky problem, therefore, we have exactly the same situation concerning existence and uniqueness of solution as for Stokes' problem: Δg must satisfy three conditions, which may be expressed in the form, analogous to (50-4),

$$\iint_{\sigma} \Delta g'(\theta, \lambda) Y_1(\theta, \lambda) d\sigma = 0, \quad (50-11)$$

which means that the upward continuation of Δg to S' must not contain any first-degree spherical harmonic.¹

Corresponding to these *three* conditions, the solution for T' , and consequently also for T , will contain *three* free constants (this is true if the linear boundary value problem is considered in itself; for physical reasons, two of these constants must be zero). Again we get a unique solution by requiring the spatial function T to have a form that contains no first-degree spherical harmonics.

The linear Molodensky problem. The general linear Molodensky problem for an arbitrary reference surface for an arbitrary reference potential, as formulated in sec. 41, is an *oblique-derivative problem*.

The classical boundary value problems--the Dirichlet problem and problems involving normal derivatives--can be formulated in terms of Fredholm integral equations of the second kind, and the well-known *Fredholm alternative* holds (cf. Kellogg, 1929, p.298):

If the homogeneous boundary-value problem has no non-zero solution, then the corresponding nonhomogeneous problem is solvable for arbitrary continuous boundary values. If the homogeneous problem has n independent solutions, then the boundary values must satisfy n independent conditions for the corresponding nonhomogeneous problem to be solvable, and the solution depends on n free parameters (because of the n independent solutions of the homogeneous problem).

An example is furnished by Stokes' problem, in which $n = 3$.

An analogous formulation of the oblique-derivative problem leads to singular integral equations for which the Fredholm alternative is, in general, no longer valid. An example is Molodensky's integral equation which is no longer a Fredholm equation of the second kind.

¹As a good approximation, $\Delta g'$ in (50-11) may be replaced by the given gravity anomaly Δg on the telluroid Σ , which means that also Δg must not contain spherical harmonics of first degree.

However, if the oblique-derivative problem is regular, that is, if the direction of the derivative is nowhere tangential to the boundary surface, then the Fredholm alternative is still valid, in spite of the singularity of the corresponding integral equation; cf. (Miranda, 1970, p.86); this means that the number of conditions on the boundary data f , given by (41-57), is equal to the degree of freedom in the solution, say n .

In the simple Molodensky problem we again had $n = 3$. In the present general linear case, n must be at least three, in view of the three degrees of freedom in the spatial shift of the origin, but perhaps there is $n = 4$ or 5 ?

Hörmander proved that even in the general form of the linear Molodensky problem n equals 3. First, the boundary-value problem defined by the boundary condition (41-30) is reformulated as follows: to determine a function T defined outside and on the telluroid Σ and satisfying the following three conditions,

1. *Harmonicity*: $\Delta T = 0$ outside Σ , (50-12)

2. *Boundary condition*: $T + \underline{m}^T \text{grad } T = f$ on Σ , (50-13)

3. *No first-degree harmonic*:

$$T(x) = \frac{c}{r} + O\left(\frac{1}{r^3}\right) \text{ as } r \rightarrow \infty, \quad (50-14)$$

c being some constant and $O\left(\frac{1}{r^3}\right)$ denoting terms of order $1/r^3$ and higher.

Hörmander proved that the corresponding homogeneous problem, that is, (50-12), (50-13), and (50-14) with $f \equiv 0$, has the unique solution $T \equiv 0$. The general solution of the homogeneous problem without imposing (50-14), therefore, contains three independent spherical harmonics of degree 1. This proves that $n = 3$ also for the general linearized Molodensky problem; in fact, if n were > 3 , then the solution of (50-12), (50-13), and (50-14) would no longer be unique.

Hörmander's proof is extremely involved and laborious and cannot be given here. Even his uniqueness theorem (Hörmander, 1976, sec.1.5) is so complicated, containing many expressions and parameters, that it cannot be stated here in full.

Let it be sufficient to mention that Hörmander's theorem contains a number of parameters which depend on properties of the earth's topography. Larger slopes of the terrain (say 60°) are permitted provided they do not occur too frequently. Although a detailed study of fitting Hörmander's para-

meters to the actual earth's topography has not yet been made, it appears that the theorem is general enough to ensure the *uniqueness* of solution of the linear Molodensky problem, with an ellipsoidal reference field, for the actual topography of the earth, the *existence* of a solution being generally guaranteed by the theory of the oblique derivative problem.

51. HÖRMANDER'S RESULTS FOR THE NONLINEAR PROBLEM

Inverse function problems. In sec. 40 we have seen that the nonlinear Molodensky problem can be regarded mathematically as an inverse function problem.

The general inverse function problem may be formulated as follows. Consider a function f such that

$$y = f(x) . \quad (51-1)$$

It can be an ordinary real function of one variable, as illustrated in Fig. 51.1, or a nonlinear operator mapping one Banach space X into another Banach space Y , so that

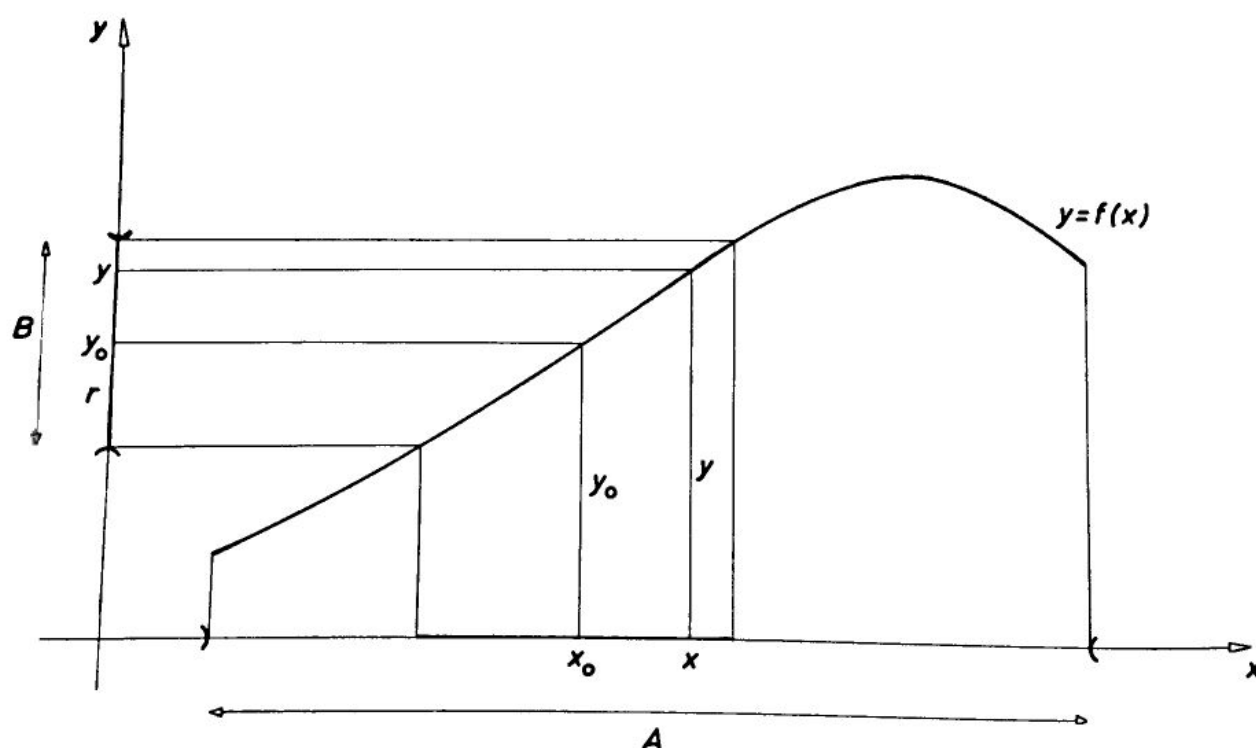


FIGURE 51.1. *The inverse-function problem.*

$$f : X \rightarrow Y, \quad (51-2)$$

x and y denoting points in the spaces X and Y , respectively. More precisely we shall assume that f is defined on an open set A of X .

Let now $x_0 \in A$ and $y_0 = f(x_0)$ be a point of Y . Suppose we can find an open neighborhood of y_0 , say the "ball" B :

$$\|y - y_0\| < r, \quad (51-3)$$

such that for each $y \in B$ there is a unique $x \in A$ for which $y = f(x)$. Then we say that f is locally invertible near y_0 , or that the inverse function f^{-1} exists:

$$x = f^{-1}(y) \quad \text{if} \quad \|y - y_0\| < r. \quad (51-4)$$

If this is possible, then we say that an *inverse function theorem* holds.

A sufficient condition is that the function f is continuously differentiable in A and that the derivative has a bounded inverse at the point x_0 . This is the inverse function theorem in its usual "elementary" form. See, e.g., (Dieudonné, 1960, sec. 10.2), (Loomis and Sternberg, 1968, sec. 3.11), or (Schwartz, 1969, p. 15); the inverse function problem is a special case of the implicit function problem as we have seen in sec. 40. The derivative, or differential, of a function in a Banach space is the so-called Fréchet derivative; it is a linear operator.

The condition of this inverse function theorem sounds natural and easy to fulfill, but it is not satisfied in many important applications. In particular, it is not satisfied in Molodensky's problem.

In this problem, the Fréchet derivative is represented by the linearization of sec. 41, and its inverse is given by the solution of the linearized Molodensky problem. This inverse is not bounded because of the "roughening effect" of differentiation; cf. p. 437.

In certain cases, the inverse function may exist even if the Fréchet derivative does not have a bounded inverse. We then speak of *advanced inverse function theorems*; they are usually hard to prove.

The usual approach to advanced implicit and inverse function theorems is by a modified Newton iteration method, as mentioned in sec. 40; cf. (Schwartz, 1969, chapter II; Sternberg, 1969; Berger, 1977, sec. 3.4).

Nash-Hörmander iteration. The treatment of the nonlinear Molodensky problem by Hörmander (1976) involves an extremely difficult inverse function

theorem. He proved it by a discrete version of a continuous method devised by J. Nash in 1956.

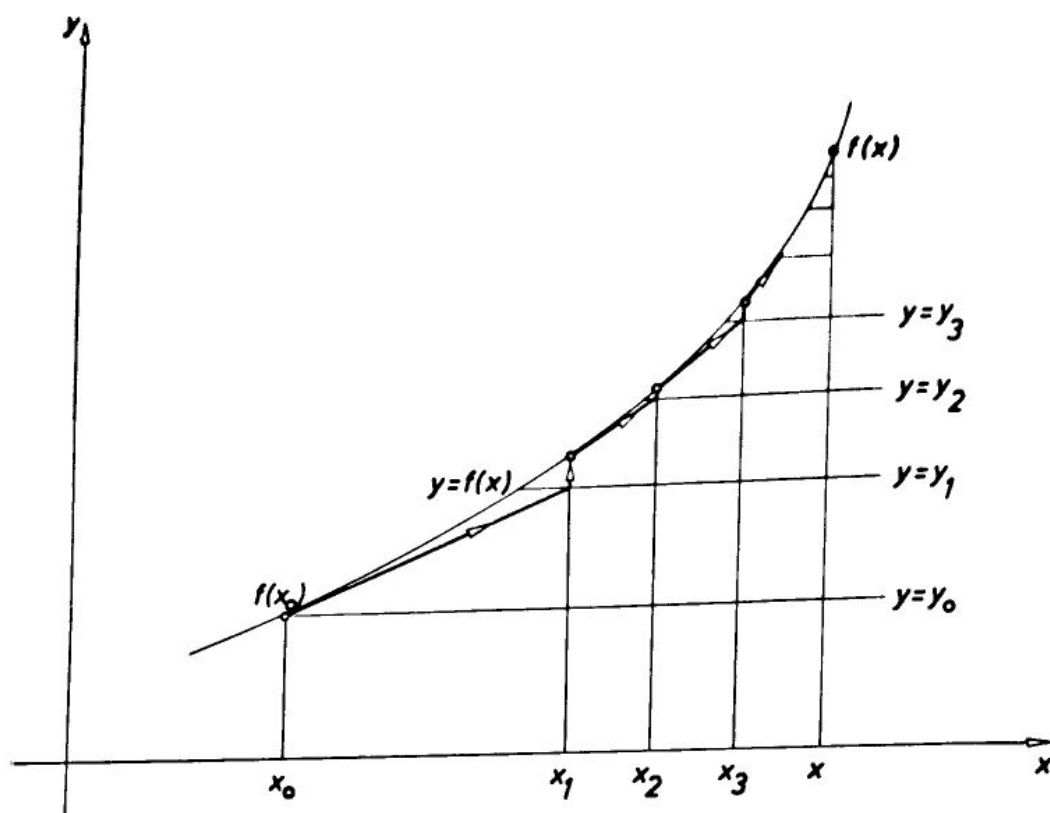


FIGURE 51.2. Nash-Hörmander iteration.

The principle of the Nash-Hörmander iteration is illustrated by Fig. 51.2. Let a value $f(x)$ be given but the corresponding x be unknown. The value x is determined in the following way. Select a point x_0 and compute the corresponding $f(x_0)$. Take a sequence y_0, y_1, y_2, \dots of points in the space Y such that

$$y_0 = f(x_0), \quad \lim_{n \rightarrow \infty} y_n = f(x). \quad (51-5)$$

Using the geometry of Fig. 51.2, the tangent at $f(x_0)$ is intersected with the line $y = y_0$, which gives x_1 . The tangent at $f(x_1)$ is now intersected with $y = y_2$, which gives x_2 , and so on. Repeating the procedure we approach the desired x as closely as we wish: there is

$$x = \lim_{n \rightarrow \infty} x_n. \quad (51-6)$$

It is instructive to compare this method with Newton's iteration as shown in Fig. 40.1 (p.336). In Newton's method, the tangents at the successive approximation points are all intersected with one and the same horizontal line through the end point. In Hörmander's method, however, the tangents at the successive approximation points x_0, x_1, x_2, \dots are intersected with different horizontals $y=y_1, y=y_2, y=y_3, \dots$ approaching the end point $f(x)$. The respective iteration procedures are illustrated in both figures by heavy lines with arrows. It is seen that the approximating line in the Nash-Hörmander method stays, so to speak, closer to the curve.

In this original simple form, the scheme of Fig. 51.2 cannot unfortunately be proved to converge for Molodensky's problem. As in other advanced inverse function problems treated by Newton's method, a suitable smoothing must be applied.

The reason for introducing a smoothing in Molodensky's problem is the following. It is a well-known difficulty with many higher order solutions that the higher order terms are getting rougher and rougher. This is the case if an iteration involves differentiation: the derivative is almost always less smooth than the original function. We have already met with this difficulty in Molodensky's series. As we have seen, e.g., in sec. 45, the calculation of higher order terms involves successive applications of an operator L which acts similarly to differentiation and is responsible for the increasing roughness of the higher terms.

Thus, a "roughening effect" is found already in the series solution of the *linear* Molodensky problem. A similar effect, due to differentiation, occurs in the iterative solution of the *nonlinear* Molodensky problem as treated by Hörmander: the functions involved get rougher and rougher, and the iteration is likely to "blow up".

More precisely, the isozenithal vector field \underline{m} as given by (41-29) involves twice differentiating the reference potential U , and this procedure is repeated at each iteration since, as we shall see below, the potential W obtained at each step serves as a reference potential U for the next step. This loss of derivatives of two orders at each step of iteration is the reason why the Fréchet derivative, given by the linearized Molodensky problem, does not have a bounded inverse in the Banach spaces used for studying the nonlinear Molodensky problem.¹

¹The reason for this and similar difficulties in Molodensky's problem is that it is a *free* boundary-value problem (the boundary surface S is "free", that is, unknown). If the earth's surface S could be considered known (in principle, it can be determined by a combination of geometrical satellite and terrestrial techniques), then the determination of the external gravitational field would lead to a much simpler *fixed* boundary-value problem; cf. (Koch, 1971; Koch and Pope, 1972).

Sansó's artifice discussed in the next sections consists precisely in transforming Molodensky's problem to a fixed boundary-value problem (in "gravity space"), which leads to an elementary inverse function problem.

Therefore, we must counteract this "roughening effect" by a suitable smoothing, taking care, however, that the degree of smoothing is successively reduced so as, in the limit, to obtain the right result. For the very complicated details we refer the reader to (Hörmander, 1976, chapter II).

The inverse, or implicit, function theorem obtained in this way is now applied to Molodensky's problem as indicated in sec. 40. The crucial problem is to prove that Molodensky's problem meets the conditions of the present implicit function theorem. Again, the reader is referred for details to (Hörmander, 1976, chapter III). We must here restrict ourselves to describing and explaining some basic ideas and the principal results.

Hölder norms. First we must say a few words about the Banach spaces and norms used in Hörmander's treatment of Molodensky's problem.

In sec. 5 we have considered the space C consisting of real-valued continuous functions f defined on a compact set B in R^n with norm

$$\|f\|_0 = \max_{\underline{x} \in B} |f(\underline{x})|, \quad (51-7)$$

\underline{x} being an abbreviation of $[x_1, x_2, \dots, x_n]$. This function space C will be denoted in the present context by H^0 .

An appropriate norm for functions that are continuously differentiable as well as continuous is

$$\|f\|_1 = \max_{\underline{x} \in B} |f(\underline{x})| + \max_{\underline{x} \in B} \left| \frac{\partial f}{\partial x_i} \right|, \quad (51-8)$$

where $\partial f / \partial x_i$ denotes any partial derivative. The functions with finite norm (51-8) form a Banach space H^1 .

Now it is of basic importance for Molodensky's problem, as for many problems in potential theory, to define norms $\|f\|_\alpha$ for $0 < \alpha < 1$, that is, spaces H^α intermediate between H^0 and H^1 . For this purpose we consider continuous functions that satisfy a Hölder condition with exponent α ; they are functions for which

$$\sup_{\underline{x}, \underline{y} \in B} \frac{|f(\underline{x}) - f(\underline{y})|}{|\underline{x} - \underline{y}|^\alpha} \quad (51-9)$$

is finite, $|\underline{x} - \underline{y}|$ denoting the distance between the points \underline{x} and \underline{y} . These functions form a Banach space H^α ; the norm is given by

$$\|f\|_{\alpha} = \max_{\underline{x} \in B} |f(\underline{x})| + \sup_{\underline{x}, \underline{y} \in B} \frac{|f(\underline{x}) - f(\underline{y})|}{|\underline{x} - \underline{y}|^{\alpha}}. \quad (51-10)$$

This norm has already been mentioned in sec. 47, cf. eq. (47-57).

It can be shown that

$$H^0 \supset H^{\alpha_1} \supset H^{\alpha_2} \supset H^1 \quad (51-11)$$

for

$$0 < \alpha_1 < \alpha_2 < 1; \quad (51-12)$$

that is, there are more functions in H^0 than in H^{α_1} , more functions in H^{α_1} than in H^{α_2} , and more functions in H^{α_2} than in H^1 . So satisfying a Hölder condition with exponent α is a stronger condition than mere continuity and weaker than differentiability.

We may also consider a Hölder condition (51-9) with exponent $\alpha = 1$; it is seen that this is almost (although not completely) the same as differentiability. In fact, we shall use H^1 for the space of functions satisfying a Hölder condition with $\alpha = 1$ rather than for functions with finite norm (51-8).

So far, we have defined spaces H^{α} for $0 \leq \alpha \leq 1$. For $\alpha > 1$ we proceed as follows. Let k be a positive integer such that $k < \alpha \leq k+1$ (for instance, for $\alpha = 5.75$ there is $k = 5$). Denote by $D^k f$ any derivative of k -th order (for instance,

$$\frac{\partial^5 f}{\partial^2 x_1 \partial^3 x_2}$$

for $k = 5$). Then the norm $\|f\|_{\alpha}$ is defined as

$$\|f\|_{\alpha} = \max_{\underline{x} \in B} |f(\underline{x})| + \sup_{\underline{x}, \underline{y} \in B} \frac{|D^k f(\underline{x}) - D^k f(\underline{y})|}{|\underline{x} - \underline{y}|^{\alpha-k}}. \quad (51-13)$$

It is clear that (51-10) is a special case of (51-13) with $k = 0$.

In other words, the space H^{α} consists of continuous functions which are k times differentiable and whose k -th derivatives satisfy a Hölder condition with exponent $\alpha - k \leq 1$.

So far we have supposed f to be a real-valued function. If f is a vector-valued function of m components f_1 :

$$f = [f_1, f_2, \dots, f_m] , \quad (51-14)$$

then its norm will simply be defined as the sum of the norms of its components:

$$\|f\|_\alpha = \|f_1\|_\alpha + \|f_2\|_\alpha + \dots + \|f_m\|_\alpha . \quad (51-15)$$

Reformulation of Molodensky's problem. Let us first recapitulate the basic assumptions for Molodensky's problem:

- (1) The earth is a rigid body rotating with a given constant angular velocity ω around an axis which is fixed with respect to the earth and which we choose as the x_3 -axis.
- (2) The axis of rotation passes through the center of gravity.
- (3) The earth's surface S is a differentiable one-to-one image of the unit sphere in a topological sense.
- (4) The earth's gravitational field does not change with time and is harmonic outside S .
- (5) The gravity vector \underline{g} and the gravity potential W are known at every point of S .

The problem consists in determining the surface of the earth and its external gravitational potential.

Hörmander has reformulated this problem in a way that a unique solution is possible. We shall employ the usual notation

$$\underline{x} = [x_1, x_2, x_3] , \quad \underline{g} = [g_1, g_2, g_3] \quad (51-16)$$

for the position vector and the gravity vector, respectively. The boundary values of the gravity vector \underline{g} and the potential W on S are denoted by $\underline{\bar{g}}$ and \bar{W} . Restriction of a spatial function f to the surface S is expressed by $f \circ S$; thus $f \circ S$ is a function defined on S .

Then Hörmander's reformulation of Molodensky's problem reads: to determine a closed surface S in R^3 , which is a one-to-one image of the unit sphere, from given values $\underline{\bar{g}}$ and \bar{W} , such that the following conditions are satisfied:

$$\bar{W} = W \circ S + \sum_{j=1}^3 a_j A_j , \quad (51-17)$$

$$\bar{g} = g \circ S = (\text{grad } W) \circ S, \quad (51-18)$$

$$W(\underline{x}) = V(\underline{x}) + \frac{1}{2} \omega^2 (x_1^2 + x_2^2), \quad (51-19)$$

$$\Delta V = 0 \quad \text{outside } S, \quad (51-20)$$

$$V(\underline{x}) = \frac{\text{const.}}{|\underline{x}|} + O\left(\frac{1}{|\underline{x}|^3}\right). \quad (51-21)$$

Uniqueness of the solution is achieved by postulating that the harmonic function $V(\underline{x})$, which represents the external gravitational potential, contains no first-degree spherical harmonics; this is expressed by (51-21),

$$|\underline{x}| = r \quad (51-22)$$

denoting the radius vector. The second term on the right-hand side of (51-19) designates, of course, the centrifugal force potential, ω being the angular velocity of the earth's rotation.

The new feature is (51-17) instead of simply taking $\bar{W} = W \circ S$, which would be the restriction of W to S . In the modified expression, the a_j are three constants to be determined, and the A_j are three suitably assumed functions. The purpose of adding the linear combination $\sum a_j A_j$ is to ensure the solvability of Molodensky's problem for arbitrary boundary data \bar{W} and \bar{g} .

This is to be understood as follows. Assume the earth's surface S to be known, and consider the given function $\bar{W} = W \circ S$. We can now solve the exterior Dirichlet problem

$$\Delta V = 0, \quad V \rightarrow \infty \quad \text{for} \quad r \rightarrow \infty, \\ V \circ S = \bar{W} - \frac{1}{2} \omega^2 (\bar{x}_1^2 + \bar{x}_2^2), \quad (51-23)$$

where \bar{x}_i represents the cartesian coordinates of a surface point. This gives, for every data function \bar{W} , a unique solution $V(\underline{x})$. For arbitrary data \bar{W} , the spatial function $V(\underline{x})$ will, in general, contain spherical harmonics of first degree, contrary to the condition (51-21).

Let now (51-17) be used instead of $\bar{W} = W \circ S$. Then (51-23) is replaced by

$$V \circ S = \bar{W} - \frac{1}{2} \omega^2 (\bar{x}_1^2 + \bar{x}_2^2) - \sum_1^3 a_j A_j. \quad (51-24)$$

Again, the exterior problem with boundary data $v \circ S$ has a unique solution $V(\underline{x})$ which will contain three linearly independent spherical harmonics of the first degree. Now, however, the three constants a_j can be chosen in such a way that these three first-degree harmonics vanish.

This is readily seen to lead to three linear equations for the three unknowns a_1, a_2, a_3 . These equations will have a unique solution, provided the boundary-value problem

$$\Delta v = 0, \quad v \circ S = \sum_{j=1}^3 a_j A_j \quad (51-25)$$

has a solution $v(\underline{x})$ which contains three linearly independent first-degree harmonics

$$(c_1 \sin \theta \cos \lambda + c_2 \sin \theta \sin \lambda + c_3 \cos \theta) / r^2. \quad (51-26)$$

In fact, c_1 can then be chosen equal to the c_1 of the boundary-value problem (51-23), and similar for c_2 and c_3 . Since the problem (51-24) is the difference of problems (51-23) and (51-25), all first-degree harmonics will cancel in the solution of (51-24).

To achieve this, we may select the functions A_i as follows:

$$A_1 = \left(\frac{\sin \theta \cos \lambda}{r^2} \right) \circ \Sigma, \quad A_2 = \left(\frac{\sin \theta \sin \lambda}{r^2} \right) \circ \Sigma, \quad A_3 = \left(\frac{\cos \theta}{r^2} \right) \circ \Sigma. \quad (51-27)$$

If the earth's surface S coincided with the telluroid Σ , then the spatial functions corresponding to the A_i by a solution of the exterior Dirichlet problem would be the first-degree harmonics

$$\frac{\sin \theta \cos \lambda}{r^2}, \quad \frac{\sin \theta \sin \lambda}{r^2}, \quad \frac{\cos \theta}{r^2} \quad (51-28)$$

themselves, so that the condition that v contains (51-26) is certainly satisfied (with $c_i = a_i$). If S does not deviate too much from Σ , then this condition is still satisfied because of continuity, even if the functions A_i are now regarded as functions on S instead of Σ by associating to a point P on S the same function value as to the corresponding point Q on Σ (cf. Fig. 41.1); of course, there will now in general be $c_i \neq a_i$.

The linearization of the modified Molodensky problem defined by (51-17) through (51-21) is now done as in sec. 41. The only difference is the additional term $\sum_1^3 a_i A_i$, so that the boundary condition (41-30) is now replaced by

$$T + \underline{m}^T \text{grad } T = \Delta W + \underline{m}^T \Delta \underline{g} - \sum_1^3 \delta_i A_i, \quad (51-29)$$

where δ_i is the difference: a_i for S minus a_i for Σ ; for the above choice (51-27) we have $\delta_i = a_i - c_i$. Clearly, the three numbers $\delta_1, \delta_2, \delta_3$ are small together with T .

As we have seen in sec. 50, the linear boundary value problem with boundary condition (51-29) for a given right-hand side f has a unique solution provided f satisfies three conditions. If the three parameters $\delta_1, \delta_2, \delta_3$ are not fixed beforehand but if we let them vary, they may be determined precisely in such a way that the three conditions are satisfied. Thus the linear boundary-value problem with boundary condition (51-29) always has a unique solution.

Outline of the iteration procedure. Without going into mathematical details, we shall now describe in broad outline how the Nash-Hörmander iteration illustrated by Fig. 51.2 is to be understood physically in its application to Molodensky's problem.

We have assumed that the function $y = f(x)$ is known; this means that, given any x_n , we are able to compute the corresponding $y_n = f(x_n)$ in a uniquely defined manner.

What are the variables x and y in Molodensky's problem? In the simplified introductory presentation given in sec. 40 we have taken $x = S$ and $y = \underline{g}$ (the gravity vector on S), but now it is appropriate to choose x and y in a somewhat different, though more complicated, way. We identify x with the triple (W, S, \underline{a}) and y with the pair $(\bar{W}, \underline{\bar{g}})$:

$$x = (W, S, \underline{a}), \quad (51-30)$$

$$y = (\bar{W}, \underline{\bar{g}}). \quad (51-31)$$

Here W is the gravity potential considered as a spatial function as usual, \underline{a} denotes the vector

$$\underline{a} = [a_1, a_2, a_3], \quad (51-32)$$

a_i being arbitrary real numbers. The function W satisfies (51-19) through

(51-21) but is arbitrary otherwise. The surface functions \bar{W} and \bar{g} are determined by (51-17) and (51-18), the A_i being given functions on S .

Given x , that is, the spatial potential W , the surface S , and the vector \underline{a} , it is therefore possible uniquely to compute \bar{W} and \bar{g} , that is, the vector y , by (51-17) and (51-18). These two equations thus represent the functional relationship $y = f(x)$.

Molodensky's problem as formulated above is the inverse problem: given \bar{W} and \bar{g} on S , to determine W , S , and \underline{a} ; or briefly, given y , to determine x .

The iteration for solving this problem may now be described in the following way. We assume a starting approximation

$$x_0 = (U, \Sigma, 0). \quad (51-33)$$

We have put

$$W_0 = U, \quad (51-34)$$

which is the normal ellipsoidal potential function, or *reference potential*, the coordinate axes being axes of the ellipsoid. We have further put

$$S_0 = \Sigma, \quad (51-35)$$

which is the telluroid; any surface approximating the earth's surface S and permitting a unique correspondence between points Q of Σ and P of S may be used, cf. sec. 41. We must only require the telluroid Σ to be a smooth surface in the sense that it is arbitrarily often differentiable.

For the vector \underline{a} we may for simplicity take $\underline{a}_0 = 0$.

We now compute $y_0 = f(x_0)$, that is we determine \bar{W}_0 and \bar{g}_0 on S_0 , which are nothing else than the normal potential U_Q and the normal gravity vector \underline{y}_Q on the telluroid Σ .

Next we apply the linearization described in sec. 41; this corresponds graphically to the tangent at the point $f(x_0)$ in Fig. 51.2. The intersection with the line $y = y_1$ represents the solution of the linearized boundary-value problem with the boundary condition (51-29) which, as we have seen above, always has a unique solution.

What is T in the present case? If the horizontal line $y = \text{const.}$ to be intersected passed through the point $y = f(x)$ for which y is given (the end point of the iteration), as it does in Newton's method (Fig. 40.1), then there would be $T = W - U$ as usual, since the given W corresponds

to the end point of the iteration just as U corresponds to the starting point. In the present case, however, the horizontal line to be intersected, $y = y_1$, does not pass through the end point. By (51-31) we have

$$y_1 = (\bar{W}_1, \bar{g}_1) ; \quad (51-36)$$

y_1 can be taken in a suitable way intermediate between the initial value

$$y_0 = (U, \bar{Y}) , \quad (51-37)$$

U and \bar{Y} denoting the normal potential and the normal gravity vector on the telluroid Σ , and the data value

$$y = (\bar{W}, \bar{g}) , \quad (51-38)$$

comprising potential and gravity vector as given on the earth's surface S . In agreement with this we then have

$$\Delta W = \bar{W}_1 - U , \quad (51-39)$$

$$\Delta \bar{g} = \bar{g}_1 - \bar{Y} . \quad (51-40)$$

The solution of the linear boundary-value problem with boundary condition (51-29) then gives a function T which we shall call T_1 . It also gives a vector $\underline{\delta}$ by (41-27), which we shall call $\underline{\delta}_1$, and a vector $\underline{\delta} = [\delta_1, \delta_2, \delta_3]$, which we shall call $\underline{\delta}_1$. We define a new reference potential U_1 by

$$U_1 = U + T_1 \quad (51-41)$$

and determine a new telluroid Σ_1 as the locus of points Q_1 defined by

$$\vec{QQ}_1 = \underline{\delta}_1 . \quad (51-42)$$

Similarly, an approximation to the vector $\underline{a} = [a_1, a_2, a_3]$ corresponding to (51-17) is obtained by

$$\underline{a}_1 = \underline{a}_0 + \underline{\delta}_1 = \underline{\delta}_1 . \quad (51-43)$$

Now we know

$$x_1 = (U_1, \Sigma_1, a_1) , \quad (51-44)$$

and we can compute

$$f(x_1) = (\bar{U}_1, \bar{\Sigma}_1) \quad (51-45)$$

by (51-17) and (51-18); note that $f(x_1) \neq y_1$ in agreement with Fig. 51.2.

Now this procedure is repeated. We take a suitable

$$y_2 = (\bar{W}_2, \bar{g}_2) \quad (51-46)$$

intermediate between y_1 and y , compute

$$\Delta W = \bar{W}_2 - \bar{U}_1 , \quad (51-47)$$

$$\Delta g = \bar{g}_2 - \bar{\Sigma}_1 , \quad (51-48)$$

and solve the corresponding linear boundary-value problem, obtaining T_2 , $\underline{\Sigma}_2$, and $\underline{\delta}_2$. We form

$$U_2 = U_1 + T_2 , \quad (51-49)$$

determine a new telluroid Σ_2 as the locus of points Q_2 such that

$$\overrightarrow{Q_1 Q_2} = \underline{\Sigma}_2 , \quad (51-50)$$

and a vector¹ \underline{a}_2 by

$$\underline{a}_2 = \underline{a}_1 + \underline{\delta}_1 . \quad (51-51)$$

Putting

$$x_2 = (U_2, \Sigma_2, \underline{a}_2) \quad (51-52)$$

¹The same functions A_i as defined by (51-27) may be used throughout the iteration; their definition is extended from Σ to all surfaces Σ_n by postulating that the value of the functions A_i is the same at all corresponding points Q, Q_1, Q_2, \dots .

we can compute $f(x_2) \neq y_2$. The procedure can be repeated arbitrarily often. The essential feature is that the spatial potential computed at a certain step is used as a reference potential for the next step, and similarly for the boundary surface.

The precise mathematical procedure also involves an appropriate smoothing as we have mentioned. Convergence can be proved using various Hölder norms and estimates for the linear problem and also for the nonlinearity. All this is exceedingly difficult and laborious. Finally one obtains Hörmander's Theorem on existence and uniqueness of Molodensky's problem:

Assume any $\epsilon > 0$, then:

(1) For all \bar{W} and \bar{g} in a $H^{2+\epsilon}$ neighborhood of \bar{W}_0 and \bar{g}_0 , the modified Molodensky problem defined by equations (51-17) to (51-21) has a solution S close to $S_0 = \Sigma$ in $H^{2+\epsilon}$ and $[a_1, a_2, a_3]$ close to 0 in R^3 .

(2) If \bar{W} and \bar{g} are in H^α for some $\alpha > 2 + \epsilon$ which is not an integer, then $S \in H^\alpha$.

(3) One can find a $H^{3+\epsilon}$ neighborhood of S_0 which cannot contain two solutions of the problem.

Let us look at this theorem more closely and explain its meaning. A $H^{2+\epsilon}$ neighborhood of \bar{W}_0 consists of all functions \bar{W} for which

$$\|\bar{W} - \bar{W}_0\|_{2+\epsilon} < \delta, \quad (51-53)$$

where δ is sufficiently small and the norm is defined by (51-13) with $a = 2 + \epsilon$. Smallness of this norm implies that not only the maximum deviation of \bar{W} from \bar{W}_0 ,

$$\max |\bar{W} - \bar{W}_0|$$

is small, but also that

$$\max |D\bar{W} - D\bar{W}_0|$$

and

$$\max |D^2\bar{W} - D^2\bar{W}_0|$$

are small, so that not only \bar{W} must be close to \bar{W}_0 , but also the first and the second derivatives of \bar{W} must be close to those of \bar{W}_0 . In

addition to this, something more is required. If $\epsilon = 1$, then also closeness of the third derivatives must hold; if $0 < \epsilon < 1$, then the intermediate Hölder condition is stronger than mere closeness of the second and weaker than closeness of the third derivatives: the difference of the second derivatives must satisfy a Hölder condition.

Closeness of the telluroid Σ and the earth's surface S in $H^{2+\epsilon}$ means that the maximum deviation of the surface is small and that, in addition, slopes (first derivatives) and curvatures (second derivatives) are also similar for S and Σ ; in addition, there is a Hölder condition for the difference of the second derivatives.

Part (1) of Hörmander's theorem asserts the *existence* of a solution provided we have a good approximation Σ for the earth's surface S not only with respect to the maximum deviation between S and Σ , but also with respect to first and second derivatives (plus a Hölder condition), and also a good approximation to potential and gravity.

This condition is obviously very restrictive. If one uses an ellipsoidal reference field and the telluroid according to the usual definition, then the actual gravity field and the earth's surface can be expected to fall short of this condition. It is, however, not required that the *initial* approximations for S and W (corresponding to x_0) satisfy this condition; it would be sufficient if any intermediate approximation x_n would meet it, because then this intermediate approximation could be considered as the starting point. Still it seems that even so the actual earth is not smooth enough to satisfy the requirements of Hörmander's theorem.

Part (2) of the theorem assures that the resulting surface S will be as smooth as the data: if the data are n times differentiable and if the n -th derivatives satisfy a Hölder condition, then the same will hold true for S .

Part (3) ensures *uniqueness* but under an even stronger condition ($H^{3+\epsilon}$ neighborhood) than for the existence theorem of Part (1) ($H^{2+\epsilon}$ neighborhood). However, Hörmander thinks it highly probable that $H^{3+\epsilon}$ could be replaced by $H^{2+\epsilon}$, so that the condition for uniqueness would be the same as for existence.

In Part (2), integer values of ϵ are excluded; this reflects the well-known fact that Hölder conditions with $\epsilon \neq 0$ are essential in potential-theoretical considerations. In Parts (1) and (3), also integer ϵ are admitted.

In conclusion we may say that Hörmander's theorem, although not directly applicable to the real earth, gives the first mathematically exact results on existence and uniqueness for Molodensky's problem and is thus of fundamental importance.

52. THE GRAVITY SPACE APPROACH

Recently, Sansø (1977), (1978a,b) has given a completely different approach to the nonlinear Molodensky problem. The idea is to use the three cartesian components of actual gravity, g_1, g_2, g_3 , as new curvilinear coordinates, instead of the cartesian coordinates x_1, x_2, x_3 themselves. Thus the potential W becomes a function of the g_i ,

$$W = W(\underline{g}) = W(g_1, g_2, g_3) . \quad (52-1)$$

On the physical earth's surface S , the three components g_i of the vector \underline{g} are given, as well as the potential W ; therefore the three curvilinear coordinates g_1, g_2, g_3 of each point of the surface S are known, or S is a known surface if expressed in terms of coordinates g_i .

There are two ways of interpreting g_i : either they may be considered as curvilinear coordinates in ordinary space, or as cartesian coordinates in an auxiliary space, called *gravity space*. Using the second interpretation, we may say that S becomes a known surface S_g in gravity space or

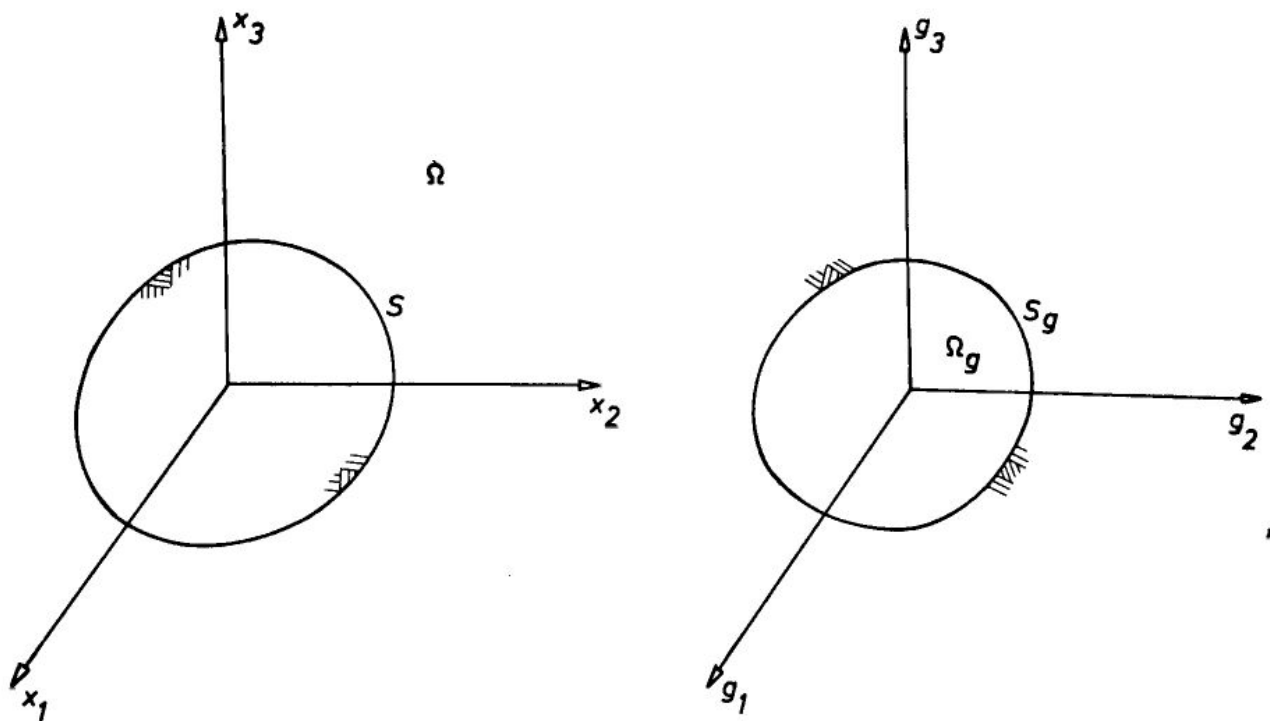


FIGURE 52.1. Ordinary space and gravity space. The surface S_g and its interior Ω_g correspond to the earth's surface S and its exterior Ω .

the free boundary-value problem is transformed into a fixed boundary-value problem; cf. Fig. 52.1. The simplification which is achieved in this way for the nonlinear Molodensky problem is so decisive that inconveniences and difficulties arising with this indirect approach are more than compensated as far as theoretical investigations on existence and uniqueness of the solution are concerned.

The main inconvenience is that a one-to-one correspondence between cartesian coordinates x_j and g_j outside and on S , which is the region in which we work, is possible only if the earth is nonrotating. To see this, consider ellipsoidal gravity γ along a radius vector in the equatorial plane. As the height increases, γ first decreases but then it increases again because the centrifugal force becomes dominant. So at a certain elevation, γ will be the same as on the ground, which violates a one-to-one correspondence between gravity vector and position. For the actual gravity field the situation is similar as in the ellipsoidal case.

If the earth is considered nonrotating, then the correspondence between gravity and position is seen to be one-to-one (provided the Marussi condition holds, see below). In other terms, the correspondence is unique if we work with the gravitational potential V and the gravitational vector $\text{grad } V$ instead of the gravity potential W and the gravity vector $\underline{g} = \text{grad } W$.

It is, of course, clear that only W and $\text{grad } W$ (including the centrifugal force) are directly measurable. However, the effect of centrifugal force can be calculated with sufficient accuracy on the basis of our current knowledge of the earth's surface (the error in the centrifugal force is less than ± 0.005 mgal for a position error of ± 10 meters), and subtracted from W and $\text{grad } W$ to give their gravitational counterparts V and $\text{grad } V$. Therefore, Sanso's boundary-value problem, which uses gravitation instead of gravity, is practically as meaningful as the original Molodensky problem.

In the sequel we shall thus work with the gravitational potential V , which is a harmonic function, and take \underline{g} as

$$\underline{g} = \text{grad } V, \quad (52-2)$$

so that \underline{g} is the vector of gravitation rather than gravity. We shall, however, continue to call \underline{g} , even if defined by (52-2), the gravity vector, to be in agreement with Sanso's terminology and with the term "gravity space" (this is consistent with current terminology if we consider the earth non-rotating).

It is clear that then the exterior Ω of S in ordinary space corresponds in a one-to-one manner to the interior Ω_g of S_g in gravity space; the infinity in ordinary space corresponds to the origin in gravity space (gravitation is zero at spatial infinity!).

We can thus reformulate Molodensky's problem in terms of V as follows: to find a function $V(\underline{x})$ which is harmonic in the exterior Ω of an unknown closed surface S ,

$$\Delta V = 0, \quad (52-3)$$

and which, together with its gradient, assumes on S the given boundary values

$$V|_S = \bar{V}(u), \quad (52-4)$$

$$(\text{grad } V)|_S = \bar{\underline{g}}(u), \quad (52-5)$$

where

$$u = [\phi, \lambda] \quad (52-6)$$

comprises astronomical latitude ϕ and longitude λ , which serve as coordinates on the surface.

We now introduce the components g_i of \underline{g} as new spatial coordinates; they are functions of the rectangular coordinates x_1, x_2, x_3 :

$$g_i = g_i(x_j). \quad (52-7)$$

If this transformation is to have an inverse,

$$x_j = x_j(g_k), \quad (52-8)$$

then the Jacobian determinant

$$\det \left[\frac{\partial g_i}{\partial x_j} \right]$$

must be nonzero everywhere on and outside S . Since

$$g_i = \frac{\partial V}{\partial x_i} , \quad (52-9)$$

this condition is

$$\det \left[\frac{\partial^2 V}{\partial x_i \partial x_j} \right] \neq 0 , \quad (52-10)$$

which is called Marussi condition (a similar condition assures the invertibility of the matrix M given by (41-20)). It will be assumed that the Marussi condition is satisfied everywhere outside and on S .

Now the potential V becomes a function of the vector \underline{g} :

$$V = V(\underline{g}) = V(g_1, g_2, g_3) . \quad (52-11)$$

As we have mentioned, this would reduce Molodensky's problem to a fixed boundary-value problem (actually a Dirichlet problem) in gravity space. Since V as a function of \underline{x} satisfies a linear partial differential equation of second order, which is Laplace's equation $\Delta V = 0$, it does the same as a function of \underline{g} since the transformation (52-7) or (52-8) transforms Laplace's equation into another linear second-order partial differential equation. However, since the transformation (52-8) is actually unknown, the coefficients of this differential equation are not known, and therefore this approach appears hopeless.

Sansô has found an ingenious way out of this difficulty by transforming not only the coordinates but also the potential, introducing an *adjoint potential*

$$\Psi = \underline{x} \cdot \underline{g} - V = x_k g_k - V ; \quad (52-12)$$

this is a *Legendre transformation* familiar from other fields (ordinary differential equations, analytical mechanics, thermodynamics, etc.).

Differentiating (52-12) with respect to g_i we get

$$\frac{\partial \Psi}{\partial g_i} = \frac{\partial x_k}{\partial g_i} g_k + x_i - \frac{\partial V}{\partial x_k} \frac{\partial x_k}{\partial g_i} = x_i$$

in view of (52-9). Thus

$$x_i = \frac{\partial \Psi}{\partial g_i} \quad (52-13)$$

or

$$\underline{x} = \text{grad}_g \Psi, \quad (52-14)$$

which shows a striking symmetry between x_i and V on the one hand and g_i and Ψ on the other hand.

Also (52-12) is completely symmetric

$$V + \Psi = x_k g_k, \quad (52-15)$$

and permits to express one potential in terms of the other:

$$\Psi = x_k \frac{\partial V}{\partial x_k} - V, \quad (52-16)$$

$$V = g_k \frac{\partial \Psi}{\partial g_k} - \Psi. \quad (52-17)$$

The matrix of second gradients of Ψ ,

$$\underline{M}_\Psi = \left[\frac{\partial^2 \Psi}{\partial g_i \partial g_j} \right] = \left[\frac{\partial x_i}{\partial g_j} \right], \quad (52-18)$$

(by (52-13)) is inverse to the matrix of second gradients of V :

$$\underline{M}_V = \left[\frac{\partial^2 V}{\partial x_i \partial x_j} \right] = \left[\frac{\partial g_i}{\partial x_j} \right], \quad (52-19)$$

cf. (41-35) and (41-36); that is,

$$\underline{M}_V = \underline{M}_\Psi^{-1}. \quad (52-20)$$

Now Laplace's operator

$$\Delta V = \frac{\partial^2 V}{\partial x_1^2} + \frac{\partial^2 V}{\partial x_2^2} + \frac{\partial^2 V}{\partial x_3^2} \quad (52-21)$$

is nothing else than the trace Tr of the matrix \underline{M}_V , and Laplace's equation may be written

$$\text{Tr} \underline{M}_V = 0 . \quad (52-22)$$

This gives us a possibility to find the corresponding partial differential equation for $\psi(g_1, g_2, g_3)$: by combining (52-20) and (52-22) we get

$$\text{Tr} (\underline{M}_\psi^{-1}) = 0 . \quad (52-23)$$

On introducing

$$\psi_{ij} = \frac{\partial^2 \psi}{\partial g_i \partial g_j} , \quad (52-24)$$

the matrix \underline{M}_ψ becomes

$$\underline{M}_\psi = \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{bmatrix} . \quad (52-25)$$

Inverting this matrix and taking the trace gives

$$\psi_{11}\psi_{22} - \psi_{12}^2 + \psi_{22}\psi_{33} - \psi_{23}^2 + \psi_{11}\psi_{33} - \psi_{13}^2 = 0 , \quad (52-26)$$

which is a partial differential equation for $\psi(g_1, g_2, g_3)$ with known coefficients (all ± 1), but unfortunately a nonlinear one.

The basic differential equation (52-26) may also be written in the form

$$(\text{Tr} \underline{M}_\psi)^2 - \text{Tr}(\underline{M}_\psi^2) = 0 , \quad (52-27)$$

which follows from the matrix identity

$$\text{Tr} \underline{A}^{-1} = [(\text{Tr} \underline{A})^2 - \text{Tr}(\underline{A}^2)] / (2 \det \underline{A})$$

and may also be verified by direct calculation.

In gravity space, the vector \underline{g} is the position vector, the components g_i serve as rectangular coordinates and gravity g serves as radius vector. In fact, we have from (40-7)

$$\begin{aligned} g_1 &= g \cos \phi \cos \Lambda, \\ g_2 &= g \cos \phi \sin \Lambda, \\ g_3 &= g \sin \phi, \end{aligned} \quad (52-28)$$

where ϕ and Λ are the astronomical coordinates (for a nonrotating earth or after removal of centrifugal effects). This shows that g, ϕ, Λ are nothing else than spherical polar coordinates in gravity space. The derivative $\partial/\partial g$ is thus a radial derivative in gravity space; we have

$$\frac{\partial \Psi}{\partial g} = \frac{\partial \Psi}{\partial g_k} \frac{\partial g_k}{\partial g} = \frac{\partial \Psi}{\partial g_k} \frac{g_k}{g},$$

using (52-28). Hence

$$g \frac{\partial \Psi}{\partial g} = g_k \frac{\partial \Psi}{\partial g_k} \quad (52-29)$$

and (52-17) may be written as

$$V = g \frac{\partial \Psi}{\partial g} - \Psi. \quad (52-30)$$

The boundary condition in gravity space thus becomes

$$\left(g \frac{\partial \Psi}{\partial g} - \Psi \right) \circ S_g = \bar{V}(u), \quad (52-31)$$

where the known function $\bar{V}(u)$ is given as a function of the parameter (52-6) which in gravity space denotes the two angular spherical coordinates; S_g is the image of the earth's surface in gravity space as in Fig. 52.1. For large values of the spatial radius vector

$$r = \sqrt{x_k x_k} = |\underline{x}| \quad (52-32)$$

we have

$$V = \frac{\mu}{r} + O\left(\frac{1}{r^3}\right), \quad (52-33)$$

$$g = \frac{\mu}{r^2} + O\left(\frac{1}{r^4}\right), \quad (52-34)$$

where

$$\mu = GM, \quad (52-35)$$

denotes the product of the gravitational constant G and the earth's mass M ; we have taken the coordinate origin at the earth's center of mass.

For $r \rightarrow \infty$ we have $g \rightarrow 0$, so that the spatial infinity corresponds to the origin in gravity space. Solving (52-34) for $1/r$,

$$\frac{1}{r} = \mu^{-\frac{1}{2}} g^{\frac{1}{2}} + O\left(g^{\frac{3}{2}}\right), \quad (52-36)$$

and substituting this into (52-33) we get

$$V = \mu^{\frac{1}{2}} g^{\frac{1}{2}} + O\left(g^{\frac{3}{2}}\right), \quad (52-37)$$

which expresses the behavior of V as $g \rightarrow 0$. Finally,

$$\Psi = -2\mu^{\frac{1}{2}} g^{\frac{1}{2}} + O\left(g^{\frac{3}{2}}\right), \quad (52-38)$$

which is verified by substitution into (52-30), taking (52-37) into account.

We thus arrive at the following formulation of the geodetic boundary-value problem in gravity space: to find the solution of the partial differential equation (52-26) in the region Ω_g inside S_g with the boundary condition (52-31) on S_g ; the earth's surface S will then be given by (52-14):

$$\underline{x} \circ S = (\text{grad}_g \Psi) \circ S_g, \quad (52-39)$$

where $\underline{x} \circ S$ denotes the position vector \underline{x} restricted to the surface S , that is, the position vector of any surface point, $\underline{x}(\phi, \lambda)$.

Since the direction $\partial/\partial g$ is the direction of the radius vector in gravity space, in general different from the normal to S_g , we have an oblique-derivative problem with a known surface S_g and a linear boundary condition (52-31), but for a nonlinear partial differential equation (52-26).

53. LINEARIZATION

The linearized equation (41-43) shows a striking formal analogy to (52-17). To take a closer look at this analogy, we shall also linearize (52-17) and other relations in gravity space. We shall follow (Moritz, 1977b, sec.8).

We shall use the concept of the *gravimetric telluroid* explained in sec. 41: there is a one-to-one correspondence between the points P of the earth's surface S and Q of the gravimetric telluroid Σ by postulating

$$\gamma_i(Q) = g_i(P), \quad (53-1)$$

that is, the normal gravity vector at Q is to be equal to the actual gravity vector at P ; cf. Fig. 41.1.

As always in the gravity space approach, we assume that the earth is nonrotating or, which is the same, that the potential is the gravitational potential V . The normal gravitational potential will be denoted by \tilde{V} . Then the disturbing potential T is

$$T = V - \tilde{V}; \quad (53-2)$$

it is the same as in the usual definition $T = W - U$ since the centrifugal potential cancels in the difference.

The adjoint potentials corresponding to V and \tilde{V} are given by (52-12):

$$\psi(g_i) = g_k x_k(g_i) - V[x_j(g_i)], \quad (53-3)$$

$$\tilde{\psi}(g_i) = g_k \xi_k(g_i) - \tilde{V}[\xi_j(g_i)]. \quad (53-4)$$

Here we have been careful in specifying the arguments. The gravity space for normal gravity is identified with the gravity space of actual gravity: equal numerical values of g_i and γ_i correspond to the same point in gravity space. It is, therefore, possible to denote the independent variable in gravity space simply by g_i , also when the normal potential is under consideration, for instance, in (53-4).

Equations (52-7) and (52-8) give the transformation between ordinary space and gravity space for actual gravity. The corresponding transformations, between ordinary space and gravity space, for normal gravity are given by

$$g_i = \gamma_i(x_j) , \quad (53-5)$$

$$x_i = \xi_i(g_j) . \quad (53-6)$$

In (53-5), g_i denote the *coordinates in gravity space*, and $\gamma_i(x_j)$ are the *functions* which express normal gravity in terms of the coordinates x_j ; the $\xi_i(g_j)$ in (53-6) are the *inverse functions* of $\gamma_i(x_j)$. This will explain the notation used in (53-3) and (53-4).

It is clear now that

$$x_i(g_j) = x_i(P) \quad (53-7)$$

are the coordinates of the point P and

$$\xi_i(g_j) = x_i(Q) \quad (53-8)$$

are the coordinates of the point Q , in view of (53-1); for the same reason, P and Q are mapped into the same point in gravity space:

$$Q_g = P_g , \quad (53-9)$$

to which both $\Psi(g_i)$ and $\tilde{\Psi}(g_i)$ in (53-3) and (53-4) refer.

Let us now calculate the difference

$$\tau = \Psi - \tilde{\Psi} , \quad (53-10)$$

which is the gravity space equivalent of the anomalous potential T as given by (53-2). Subtracting (53-3) and (53-4) we get

$$\begin{aligned} \tau(g_i) = & g_k [x_k(g_i) - \xi_k(g_i)] - \\ & - V[x_j(g_i)] + \tilde{V}[\xi_j(g_i)] . \end{aligned} \quad (53-11)$$

In agreement with (41-15) we put

$$x_j = \xi_j + \zeta_j \quad (53-12)$$

(we omit the argument g_i : x_j denotes the coordinates of P and ξ_j those of Q). Now to $V[x_j(g_i)]$ we apply Taylor's theorem:

$$\begin{aligned} V(x_j) &= V(\xi_j + \zeta_j) = V(\xi_j) + \frac{\partial V}{\partial x_k} \zeta_k \\ &= V(\xi_j) + g_k \zeta_k . \end{aligned} \quad (53-13)$$

The substitution of (53-12) and (53-13) into (53-11) gives

$$\tau(g_i) = g_k \zeta_k - V(\xi_j) - g_k \zeta_k + \tilde{V}(\xi_j) = -V(\xi_j) + \tilde{V}(\xi_j)$$

or

$$\tau(g_i) = -T[\xi_j(g_i)] . \quad (53-14)$$

In geometrical terms, τ at $P_g = Q_g$ equals the negative of T at Q .

We have thus obtained the result that the adjoint potential of T is simply the negative of T . This is certainly surprising at first sight, and it indicates a deep relation between gravity space and ordinary space: gravity space is not just an artifice introduced *ad hoc*, but a natural expression of the mathematical structure of the geodetic boundary-value problem.

This will even become more evident if we consider the boundary condition. In view of (53-1) we have

$$S_g = \Sigma_g , \quad (53-15)$$

so that the earth's surface S and the telluroid Σ are mapped into the same surface S_g in gravity space.

By (52-29), the boundary condition (52-31) becomes

$$\left(g_k \frac{\partial \psi}{\partial g_k} - \psi \right) \circ S_g = \bar{V}(u) . \quad (53-16)$$

The corresponding condition for the normal potential \tilde{V} at the telluroid Σ is

$$\left(g_k \frac{\partial \tilde{\psi}}{\partial g_k} - \tilde{\psi}\right) \circ S_g = \tilde{V}(u) . \quad (53-17)$$

The subtraction of these two equations, which are linear in ψ and $\tilde{\psi}$, gives by (53-10):

$$\left(g_k \frac{\partial \tau}{\partial g_k} - \tau\right) \circ S_g = V(u) - \tilde{V}(u) . \quad (53-18)$$

Now,

$$\bar{V}(u) - \tilde{V}(u) = \bar{W}(u) - \bar{U}(u) = W_P - U_Q$$

because \bar{W} refers to S and \bar{U} to Σ and because the difference of the centrifugal potentials at P and at Q is negligibly small. By (41-3) this is

$$V(u) - \tilde{V}(u) = \Delta W . \quad (53-19)$$

Furthermore τ on S_g equals $-T$ on Σ . Thus (53-18) becomes

$$-g_k \frac{\partial T}{\partial g_k} + T = \Delta W , \quad (53-20)$$

which is now a boundary condition on the telluroid Σ . The replacement of g_k by γ_k changes (53-20) only by second-order quantities, which are to be neglected. Thus the boundary condition on Σ finally takes the form

$$T - \gamma_k \frac{\partial T}{\partial \gamma_k} = \Delta W . \quad (53-21)$$

This is nothing else than (41-43) with $\Delta g = 0$ for the gravimetric telluroid, and with (41-44). We thus have recovered the fundamental boundary condition of sec. 41 via gravity space.

What about the differential equation which τ must satisfy? We could derive it from (52-26), but there is a much simpler way, using (53-14). In this equation we substitute (53-5) and (53-6), obtaining

$$\tau[\gamma_i(x_j)] = -T(x_i) . \quad (53-22)$$

Since T satisfies Laplace's equation

$$\Delta T = 0, \quad (53-23)$$

$\tau = -T$ will also satisfy it:

$$\Delta \tau = 0; \quad (53-24)$$

if τ is considered a function of γ_i , then the Laplacian is to be expressed in terms of γ_i , which here are to be regarded as curvilinear coordinates in ordinary space related to the x_i by (53-5). It is not difficult to transform the Laplacian to curvilinear coordinates; cf. (Hotine, 1969, p.19); the important thing to note is that it does *not* have the "cartesian" form:

$$\frac{\partial^2}{\partial \gamma_1^2} + \frac{\partial^2}{\partial \gamma_2^2} + \frac{\partial^2}{\partial \gamma_3^2}. \quad (53-25)$$

Thus, as far as the *linear* problem goes, the gravity space approach simply amounts to the use of curvilinear coordinates in ordinary space. It is, therefore, not essentially different from the usual approach outlined in sec. 41; it is even less general as it supposes a nonrotating earth. The situation is quite different for the *nonlinear* problem where the gravity space approach introduces essentially new features and a considerable simplification.

Different as the ordinary approach and the use of gravity space are, *the linearized problem is the same in both methods*. This is practically important because the linearized Molodensky problem is probably sufficient for all present applications, as we have pointed out at the end of sec. 42.

Even for the linear problem, however, the gravity space approach provides a deeper insight into the problem; in particular, the structure of the operator that acts on T in (41-43),

$$T = \gamma_i \frac{\partial T}{\partial \gamma_i}, \quad (53-26)$$

is interpreted by the relation between potential and adjoint potential as expressed by (52-17).

Spherical approximation. Let us finally introduce a spherically symmetric normal potential; this corresponds to the "spherical approximation" outlined in sec. 42.

For a spherically symmetric mass configuration we have

$$\hat{V} = \frac{\mu}{r} ;$$

cf. (52-33). By a simple change of scale of length and without loss of generality we can make $\mu = 1$, obtaining

$$\hat{V} = \frac{1}{r} . \quad (53-27)$$

Differentiation with respect to x_i gives

$$\gamma_i = - \frac{x_i}{r^3} ,$$

so that

$$\gamma = \frac{1}{r^2} \quad (53-28)$$

with

$$\gamma^2 = \gamma_k \gamma_k , \quad r^2 = x_k x_k . \quad (53-29)$$

It is, therefore, possible to express the x_i in terms of γ_i by

$$x_i = - \gamma^{-\frac{3}{2}} \gamma_i ; \quad (53-30)$$

in the case of a spherically symmetric mass configuration, cartesian coordinates x_i and gravimetric coordinates γ_i are thus related in a simple way.

Another possibility to convert gravimetric coordinates into cartesian coordinates, denoted by y_i , is by putting

$$y_i = \gamma^{-\frac{1}{2}} \gamma_i . \quad (53-31)$$

These coordinates y_i can be interpreted in the following way. Let us consider an inversion in the unit sphere $r = 1$; see sec. 6. This inversion transforms a point with coordinates x_i into a point with coordinates x'_i given by

$$x'_i = \frac{1}{r^2} x_i, \quad (53-32)$$

the inverse transformation being

$$x_i = \frac{1}{r'^2} x'_i \quad \text{with} \quad r'^2 = x'_k x'_k; \quad (53-33)$$

cf. eq. (6-15). On substituting (53-30) and comparing the result with (53-31) we see that

$$y_i = -x'_i, \quad (53-34)$$

so that, apart from the sign, y_i are the cartesian coordinates of the image of the point x_i under an inversion in the unit sphere.

The corresponding transformation of harmonic functions is the Kelvin transformation (6-17): if $U(x_i)$ is a harmonic function of x_i in a domain T , then

$$V(x'_i) = \frac{1}{r'} U\left(\frac{x'_i}{r'^2}\right) \quad (53-35)$$

is a harmonic function of x'_i in the domain T' into which T is carried by the inversion.

So far, we have interpreted this transformation as a *point transformation*, which transforms a point $P(x_i)$ into a point $P'(x'_i)$, the coordinates x_i and x'_i referring to the same cartesian coordinate system. We may, however, interpret it also as a *coordinate transformation*, by which the same point in space is referred to different coordinate systems x_i and x'_i . Then the Kelvin transformation implies that if $U(x_i)$ satisfies Laplace's equation in "cartesian form" using x_i :

$$\Delta_x U = \frac{\partial^2 U}{\partial x_1^2} + \frac{\partial^2 U}{\partial x_2^2} + \frac{\partial^2 U}{\partial x_3^2} = 0, \quad (53-36)$$

then the function (53-35) satisfies Laplace's equations in cartesian form using x'_i :

$$\Delta_{\mathbf{x}'} V = \frac{\partial^2 V}{\partial x_1'^2} + \frac{\partial^2 V}{\partial x_2'^2} + \frac{\partial^2 V}{\partial x_3'^2} = 0 . \quad (53-37)$$

In view of (53-34), Laplace's operator will then have cartesian form also in coordinates y_i :

$$\Delta_{\mathbf{y}} V = 0 . \quad (53-38)$$

The symbol $\Delta_{\mathbf{x}}$, $\Delta_{\mathbf{y}}$, etc. will be reserved for Laplace's operator in cartesian form.

Let us now apply these considerations to the present problem. We have seen that $T(x_i)$ satisfies Laplace's equation $\Delta_{\mathbf{x}} T = 0$; cf. (53-23). $\tau(\gamma_i)$ also satisfies Laplace's equation (53-24), but not in cartesian form (53-25). If in $\tau(\gamma_i)$ we introduce new coordinates y_i , defined by (53-31), putting

$$\gamma_i = y_i y \quad \text{with} \quad y^2 = y_k y_k , \quad (53-39)$$

then the new function

$$\phi(y_i) = \frac{\tau}{y} = - \frac{T}{y} \quad (53-40)$$

will satisfy

$$\Delta_{\mathbf{y}} \phi = 0 \quad (53-41)$$

because of (53-35) with $U = -T$, $V = \phi$ and $y_i = -x'_i$, since T satisfies $\Delta_{\mathbf{x}} T = 0$.

Also the function v defined by

$$v(y_i) = \frac{1}{2} \left(y_i \frac{\partial \phi}{\partial y_i} - \phi \right) \quad (53-42)$$

is harmonic:

$$\Delta_y v = 0 . \quad (53-43)$$

This can be easily verified by direct calculation: there is

$$2\Delta_y v = \Delta_y \phi + y_1 \frac{\partial}{\partial y_1} \Delta_y \phi . \quad (53-44)$$

The interpretation of v is as follows. Consider ΔW as given by (53-21). (Of course, Δ has here nothing to do with Laplace's operator!) It may also be expressed in terms of τ by

$$\Delta W = \gamma_k \frac{\partial \tau}{\partial \gamma_k} - \tau . \quad (53-45)$$

In (53-21), ΔW has been considered as defined on the telluroid Σ . It may, however, also be regarded as a spatial function, defined outside and on Σ , since τ is a function of the γ_k which can be interpreted as curvilinear coordinates in space. If now ΔW , regarded as a spatial function, is expressed in terms of y_i , we can also transform (53-45) to these coordinates. This is best done by transforming it first to the form

$$\Delta W = \gamma \frac{\partial \tau}{\partial \gamma} - \tau , \quad (53-46)$$

using (52-29) with γ_i instead of g_i . Now

$$\gamma = \sqrt{\gamma_k \gamma_k} \quad (53-47)$$

is related to

$$y = \sqrt{y_k y_k} \quad (53-48)$$

by

$$\gamma = y^2 , \quad y = \sqrt{\gamma} , \quad (53-49)$$

by (53-39). Therefore,

$$\gamma \frac{\partial \tau}{\partial \gamma} = y^2 \frac{\partial \tau}{\partial y} \frac{dy}{d\gamma} = \frac{1}{2} y \frac{\partial \tau}{\partial y} , \quad (53-50)$$

and (53-46) takes the form

$$\Delta W = \frac{1}{2} y \frac{\partial \tau}{\partial y} - \tau . \quad (53-51)$$

Substituting

$$\tau = y \phi \quad (53-52)$$

according to (53-40), we get

$$\Delta W = \frac{1}{2} y \frac{\partial (y \phi)}{\partial y} - y \phi = \frac{1}{2} y \left(y \frac{\partial \phi}{\partial y} - \phi \right) . \quad (53-53)$$

Using again (52-29) with y_k instead of g_k we obtain

$$\Delta W = \frac{1}{2} y \left(y_1 \frac{\partial \phi}{\partial y_1} - \phi \right) , \quad (53-54)$$

and the comparison with (53-42) shows that

$$v = \frac{\Delta W}{y} . \quad (53-55)$$

This furnishes the desired physical interpretation of v .

These two auxiliary functions

$$\phi(y_1) = \frac{\tau}{y} = - \frac{T}{y} , \quad (53-56)$$

$$v(y_1) = \frac{1}{2} \left(y_1 \frac{\partial \phi}{\partial y_1} - \phi \right) = \frac{\Delta W}{y} , \quad (53-57)$$

which, in the case of spherical symmetry, satisfy Laplace's equation:

$$\Delta_y \phi = 0 , \quad (53-58)$$

$$\Delta_y v = 0 , \quad (53-59)$$

will play a basic role in the next section.

54. SANSÒ'S TREATMENT OF THE NONLINEAR PROBLEM

Reformulation of the problem. Let ψ be a solution of the boundary-value problem defined by the differential equation (52-26) and the boundary condition (52-31), which, in view of (52-29), may be written in the form

$$g_k \frac{\partial \psi}{\partial g_k} - \psi = V(u) \quad \text{on } S_g. \quad (54-1)$$

Then the function

$$\hat{\psi} = \psi + c_i g_i, \quad (54-2)$$

with an arbitrary constant vector c_i , is also a solution of the problem. In fact,

$$\hat{\psi}_{ij} = \frac{\partial^2 \hat{\psi}}{\partial g_i \partial g_j} = \frac{\partial^2 \psi}{\partial g_i \partial g_j} = \psi_{ij}, \quad (54-3)$$

so that $\hat{\psi}$ satisfies (52-26) if ψ does, and

$$g_k \frac{\partial \hat{\psi}}{\partial g_k} - \hat{\psi} = g_k \frac{\partial \psi}{\partial g_k} - \psi_k, \quad (54-4)$$

so that the boundary condition is also satisfied.

It is easily seen that the addition of the term $c_i g_i$ to ψ represents a translation by the vector c_i in ordinary space: by (52-13) we get

$$\hat{x}_i = \frac{\partial \hat{\psi}}{\partial g_i} = \frac{\partial \psi}{\partial g_i} + c_i = x_i + c_i. \quad (54-5)$$

We obtain a unique solution by requesting ψ to have the form (52-38), which places the x -coordinate system at the earth's center of mass. This is in complete correspondence with the usual treatment of Molodensky's problem.

However, the solution will not exist for arbitrary boundary values \bar{V} but only for those functions $V(u)$ which satisfy n conditions; from the discussion of the linearized problem we expect $n = 3$. It is true

that if we had idealized conditions, especially absence of measuring errors, then the data function $V(u)$ would satisfy these conditions because the solution exists for physical reasons. In practice, however, especially because of measuring and interpolation errors, we cannot expect that the actual $V(u)$ will exactly satisfy these conditions.

This suggests a reformulation of the boundary-value problem in gravity space along the lines of Hörmander's formulation; cf. sec. 51, especially eq. (51-17): we replace the boundary condition (54-1) by

$$g_k \frac{\partial \psi}{\partial g_k} - \psi = V(u) + a_1 g_1 \quad \text{on } S_g. \quad (54-6)$$

The new boundary-value problem can now be expected to have a solution for arbitrary data functions $V(u)$. The three constants a_1, a_2, a_3 are determined as unknowns and, so to speak, take care of the three conditions.

Transformation of the differential equation. The main difficulty in the gravity space approach lies in the differential equation for the adjoint potential ψ . This equation, given by (52-26) or (52-27), is unfortunately considerably more complicated than Laplace's equation for the original potential V .

The consideration of the spherical approximation in the preceding section suggests, however, that it may be possible to reduce, at least approximately, this differential equation to Laplace's equation.

First, in agreement with (52-38), we split off the main part in ψ by putting

$$\psi(g_i) = -2\mu \frac{1}{2} g^{\frac{1}{2}} + \tau(g_i), \quad (54-7)$$

where

$$g^2 = g_k g_k. \quad (54-8)$$

This may be interpreted by (53-10) as using a spherically symmetric reference potential

$$\tilde{\psi} = -2\mu \frac{1}{2} g^{\frac{1}{2}} \quad (54-9)$$

in gravity space (the reader will find it best to consider all transformations to follow as *transformations in gravity space* and to forget, for the

time being, about ordinary space). In contrast to the linear treatment in the preceding section we shall not introduce any approximations, so that the transformed differential equations will be as rigorous as the original one, eq. (52-26).

The reference potential (54-9), which is spherically symmetric in gravity space, is the adjoint potential of a potential \tilde{V} that is spherically symmetric in ordinary space. In fact, by (52-13) and (52-15),

$$x_k = \frac{\partial \tilde{\Psi}}{\partial g_k} = -\mu^{\frac{1}{2}} g^{-\frac{3}{2}} g_k, \quad (54-10)$$

$$r = \sqrt{x_k x_k} = \mu^{\frac{1}{2}} g^{-\frac{1}{2}}, \quad (54-11)$$

$$\tilde{V} = g_k x_k - \tilde{\Psi} = -\mu^{\frac{1}{2}} g^{\frac{1}{2}} + 2\mu^{\frac{1}{2}} g^{\frac{1}{2}} = \mu^{\frac{1}{2}} g^{\frac{1}{2}} = \frac{\mu}{r}. \quad (54-12)$$

Eq. (53-31) suggests the substitution

$$y_i = g^{-\frac{1}{2}} g_i. \quad (54-13)$$

(Now, however, the y_i are to be considered as curvilinear coordinates in gravity space, having no direct relation with cartesian coordinates in ordinary space.) This transforms the reference potential into

$$\tilde{\Psi} = -2\mu^{\frac{1}{2}} y, \quad (54-14)$$

eliminating the singularity $g^{\frac{1}{2}}$ at the origin $g = 0$.

We now introduce the new function

$$\phi = \frac{\tau}{y}, \quad (54-15)$$

so that

$$\tau = y \phi. \quad (54-16)$$

If we neglected all squares and higher powers of τ , we should have the linear spherical approximation discussed in the preceding section since, apart from a scale factor, (54-12) is identical to (53-27). This shows that ϕ , as a function of y , must satisfy a differential equation of form

$$\Delta_y \phi = O(\phi^2) ;$$

(54-17)

there can be no term $O(\phi)$ on the right-hand side since $\Delta_y \phi = 0$ as a linear approximation, by (53-58).

In fact, Sansó (1977) has calculated the exact differential equation which ϕ must satisfy. This is done by substituting

$$\psi = -2\mu \frac{1}{2}y + y\phi$$

(54-18)

into (52-27) and performing some transformations which are too lengthy to be given here. The result is rather simple:

$$\Delta_y \phi = \mu^{-\frac{1}{2}} B_1(\phi, \phi) ,$$

(54-19)

where $B_1(\phi, \phi)$ is a quadratic operator given by

$$B_1(\phi, \phi) = \frac{1}{2}(\phi - y\phi')\Delta_y \phi + y^2[(\text{Tr } \underline{L})^2 - \text{Tr } (\underline{L}^2)]$$

(54-20)

(it must be quadratic since the original equation (52-26) is). The matrix \underline{L} has elements

$$L_{ij} = \left(\delta_{ik} - \frac{3}{4} \frac{y_i y_k}{y^2} \right) \phi_{kj}$$

(54-21)

where δ_{ij} denotes the elements of the unit matrix and

$$\phi_{ij} = \frac{\partial^2 \phi}{\partial y_i \partial y_j} ;$$

(54-22)

ϕ' is defined by

$$\phi' = \frac{\partial \phi}{\partial y} ,$$

and $\Delta_y \phi$ expresses the Laplace operator in the "cartesian form"

$$\Delta_y \phi = \frac{\partial^2 \phi}{\partial y_1^2} + \frac{\partial^2 \phi}{\partial y_2^2} + \frac{\partial^2 \phi}{\partial y_3^2} ; \quad (54-23)$$

needless to say, y_i are not rigorously to be interpreted as cartesian coordinates in ordinary space.

The boundary operator (54-1)

$$g_k \frac{\partial \psi}{\partial g_k} - \psi = g \frac{\partial \psi}{\partial g} - \psi \quad (54-24)$$

is transformed as follows. Using (54-9) we find

$$g \frac{\partial \tilde{\psi}}{\partial g} - \tilde{\psi} = \mu^{\frac{1}{2}} g^{\frac{1}{2}} , \quad (54-25)$$

so that

$$g \frac{\partial \psi}{\partial g} - \psi = \mu^{\frac{1}{2}} g^{\frac{1}{2}} + g \frac{\partial \tau}{\partial g} - \tau . \quad (54-26)$$

Since

$$y = g^{\frac{1}{2}} \quad (54-27)$$

by (54-13), we have

$$\frac{\partial \tau}{\partial g} = \frac{\partial \tau}{\partial y} \frac{dy}{dg} = \frac{1}{2} g^{-\frac{1}{2}} \frac{\partial \tau}{\partial y} . \quad (54-28)$$

In view of these relations we get

$$g_k \frac{\partial \psi}{\partial g_k} - \psi = \mu^{\frac{1}{2}} y + \frac{1}{2} y \frac{\partial \tau}{\partial y} - \tau . \quad (54-29)$$

On substituting (54-16) and taking (54-6) into account we find as boundary condition for ϕ :

$$y \frac{\partial \phi}{\partial y} - \phi = 2[\bar{V}(u) + a_i \bar{y}_i] \quad \text{on } S_g , \quad (54-30)$$

where

$$\bar{v}(u) = \frac{\bar{V}(u)}{\bar{y}(u)} - \mu^{-\frac{1}{2}} \quad (54-31)$$

is a function of the data, and

$$\bar{y}(u) = y \circ S_g, \quad \bar{y}_1(u) = y_1 \circ S_g \quad (54-32)$$

denote the values of y and y_1 calculated for that point of the surface S_g which has the parameter u .

Since the direction of $\partial/\partial y$, as well as the direction of $\partial/\partial g$, is the direction of the radius vector in gravity space, we still have an oblique derivative problem as in the original formulation given at the end of sec. 52; the problem is, however, simplified because we now have a "quasilinear" differential equation (54-19), which has a form suitable for an iterative solution.

It is possible to transform the problem still further by introducing a new potential v by

$$v = \frac{1}{2} \left(y_k \frac{\partial \phi}{\partial y_k} - \phi \right) = \frac{1}{2} \left(y \frac{\partial \phi}{\partial y} - \phi \right) . \quad (54-33)$$

This substitution has been motivated in the preceding section; cf. (53-57). As a linear approximation, $v(y_1)$ is harmonic and is, furthermore, related to the potential anomaly ΔW .

In fact, we even have rigorously

$$V(g_1) = \mu^{\frac{1}{2}} g^{\frac{1}{2}} + yv, \quad (54-34)$$

so that yv represents the perturbation in the potential V , if expressed in gravimetric coordinates g_1 , in the same way as we had

$$\psi(g_1) = -2\mu^{\frac{1}{2}} g^{\frac{1}{2}} + y\phi, \quad (54-35)$$

$\tau = y\phi$ representing the perturbation in the adjoint potential ψ . It is easy to verify (54-34) by substituting (54-35) into (52-30).

By means of (54-33), eq. (54-19) is finally transformed into a differential equation (more precisely, an integrodifferential equation) for v :

$$\Delta_y v = \mu^{-\frac{1}{2}} B_2(v, v) \quad (54-36)$$

where the quadratic operator B_2 is given by

$$B_2(v, v) = -v \Delta_y v - v' \int_0^y \Delta_y v dy + 2[(\text{Tr } \underline{N})^2 - \text{Tr}(\underline{N}^2)] \\ + 4y[\text{Tr } \underline{M} \cdot \text{Tr } \underline{N} - \text{Tr}(\underline{MN})] . \quad (54-37)$$

\underline{M} and \underline{N} are 3×3 matrices with elements

$$M_{ij} = \left(\delta_{ik} - \frac{3}{4} \frac{y_i y_k}{y^2} \right) v_{kj} , \quad (54-38)$$

$$N_{ij} = \left(\delta_{ik} - \frac{3}{4} \frac{y_i y_k}{y^2} \right) \int_0^y v_{kj} dy , \quad (54-39)$$

where δ_{ij} denotes the elements of the unit matrix and

$$v_{ij} = \frac{\partial^2 v}{\partial y_i \partial y_j} ; \quad (54-40)$$

v' is defined as

$$v' = \frac{\partial v}{\partial y} \quad (54-41)$$

and Δ_y again denotes the "cartesian form" of the Laplace operator.

The solution of (54-36) has to satisfy the boundary condition

$$v \circ S_g = \bar{v} + a_i \bar{y}_i , \quad (54-42)$$

$\bar{v} = \bar{v}(u)$ and $\bar{y} = \bar{y}(u)$ being given by (54-31) and (54-32). This is simply the *boundary condition for a Dirichlet problem*.

By means of the substitution (54-33) it has thus been possible to transform Sansô's problem into a Dirichlet problem for the nonlinear equation (54-36). The price to be paid is that this equation is a nonlinear integro-differential equation, as (54-37) shows. However, since the principal part of (54-36) is simply Laplace's equation, the quadratic right-hand side be-

ing comparatively small, our equation is still relatively manageable (it is hardly necessary to remind the reader that (54-36) is as rigorous as the original equation (52-26); no neglects are involved).

This reduction to a Dirichlet problem is similar to methods used in the linear Molodensky problem, cf. (Brovar, 1964), (Krarup, 1973: the "Prague method") and the reduction to the Brillouin sphere in the present sec. 50. The enormous advantage of the gravity space approach is that the boundary condition (54-1) is linear even for the nonlinear problem, so that methods can be used that are applicable to the Molodensky problem only in its linearized form.

A necessary condition for the existence of the solution is

$$\left(\frac{\partial v}{\partial y_i} \right)_{y=0} = 0. \quad (54-43)$$

In fact, the differentiation of (54-33) gives

$$2 \frac{\partial v}{\partial y_i} = y_k \frac{\partial^2 \phi}{\partial y_i \partial y_k}. \quad (54-44)$$

If the solution ϕ is to be regular with finite second derivatives at the origin $y = \sqrt{y_j y_j} = 0$, then (54-44) must tend to zero as $y_k \rightarrow 0$. The condition (54-43) is to be provided for by suitably disposing of the free constants a_1, a_2, a_3 in the boundary condition (54-30).

If a solution v satisfying (54-43) has been found, then ϕ is obtained by

$$\phi = -2v(0) + 2y \int_0^y [v(y) - v(0)] y^{-2} dy; \quad (54-45)$$

it is easy to verify by direct substitution that this solution satisfies (54-33). Then the adjoint potential ψ is given by (54-35), and finally the earth's surface is obtained by (52-39).

Study of existence and uniqueness. There are thus two possibilities for formulating Sansô's problem in a way suitable for an iterative solution. We may consider it either an oblique-derivative problem for the partial differential equation (54-19) with the boundary condition (54-30), or a Dirichlet problem for the integrodifferential equation (54-36) with the boundary condition (54-42).

In both formulations we have boundary-value problems with a *fixed boundary*, to which the "elementary" inverse function theorem can be applied (the reason why we had to use an advanced inverse function theorem in Molodensky's problem is that it is a *free* boundary-value problem). It is easier to verify the conditions of applicability of the inverse function theorem for the second formulation, in terms of Dirichlet's problem (Sansô, 1976, 1977); the application of Newton's method gives even definite numerical estimates. The result is that a uniform solution exists provided \bar{v} satisfies the condition

$$\|\bar{v}(u)\|_{2+\epsilon} < \delta, \quad (54-46)$$

where the constant δ is sufficiently small. The norm is a Hölder norm very similar to the norm used in sec. 51.

Since \bar{v} , as given by (54-31), has the character of an anomalous potential (actual potential minus spherical reference potential), this result is very similar to Hörmander's result (51-53). It is derived much more easily but is restricted to a nonrotating earth and a *spherical* reference potential, whereas $W_0 = U$ in (51-53) can be the usual ellipsoidal reference potential. On the other hand, closeness in $H^{2+\epsilon}$ is sufficient even for uniqueness whereas in Hörmander's theorem we had for uniqueness even to require closeness in $H^{3+\epsilon}$; cf. p. 448.

An essential progress, both with respect to his first result (54-46) and to Hörmander's result, has been achieved by Sansô (1978b) by using the first formulation in terms of the oblique-derivative problem given by (54-19) and (54-30). Again the "elementary" implicit function theorem can be applied but the conditions for its applicability are more difficult to verify. In return for this one gets a much stronger result: a unique solution exists already if

$$\|\bar{v}(u) - \bar{v}_0(u)\|_{1+\epsilon} < \delta. \quad (54-47)$$

The function $\bar{v}_0(u)$ may now be the usual *ellipsoidal* normal potential, or a similar reference potential; we are no longer restricted to a spherical reference potential. The essential improvement, however, is the replacement of the norm $\|\cdot\|_{2+\epsilon}$ by $\|\cdot\|_{1+\epsilon}$: we no longer need closeness of the second derivatives, but only closeness of v and v_0 together with its first derivatives plus a Hölder condition; cf. p. 448.

It is not known, however, whether the condition (54-47) is satisfied for the real earth since we only know that a number δ , ensuring existence and

uniqueness of the solution, exists but not how great it is. The number δ may well be so small that (54-47) does not hold for the real potential.

Concluding remarks. During the last few years, the problem of existence and uniqueness of the solution for Molodensky's problem has for the first time been treated with adequate mathematical rigor. Certainly, existence and uniqueness have been proved only under very restrictive conditions on smoothness and smallness of the deviations from a "normal" solution, conditions which we cannot expect to be met in the actual geodetic situation. However, these results have been obtained *rigorously*.

The treatment by Hörmander uses a very advanced inverse function theorem and is mathematically extremely complicated; it applies to a rotating earth. The mathematical complexity is mainly due to the fact that Molodensky's problem is a free boundary-value problem, the boundary surface being unknown.

The gravity space approach due to Sansô transforms the free boundary problem into a fixed one, although for a nonlinear partial differential equation. It nevertheless reduces essentially the mathematical complexity. The limitation of the gravity space approach is the restriction to a non-rotating earth; practically this amounts to the use of gravitation instead of gravity by reducing for the effect of centrifugal force.¹

The results obtained by Sansô are stronger: he only requires closeness of \bar{V} and \bar{V}_0 in $H^{1+\epsilon}$ for existence and uniqueness, whereas Hörmander had to require closeness in $H^{2+\epsilon}$ for existence and even closeness in $H^{3+\epsilon}$ for uniqueness.

Thus, from a theoretical point of view, the impact of the gravity space approach to the geodetic boundary-value problem appears enormous. It throws new light on this problem and provides powerful new methods for studying its mathematical aspects.

From a practical point of view, on the other hand, it is important to note that the linearization (linear in the anomalous potential T) is the same in the usual approach and in gravity space, as we have seen in sec. 53. All usual methods for practically solving Molodensky's problem, as discussed in secs. 43 through 49, are based on this linearization. The gravity space approach does not give a new contribution to methods of this type; the essential advantage of the new approach manifests itself in the non-linear problem.

¹ It may be remarked that, in a work that has not yet been published, Sansô has also treated (by an iterative procedure) the problem of a rotating earth, but even Sansô's original problem is practically meaningful as we have pointed out at the beginning of sec. 52.

55. GEODYNAMICAL EFFECTS

Throughout the present book we have assumed the following idealized situation: the earth is a rigid body which rotates with constant angular velocity around an axis which is fixed with respect to the earth and passes through the earth's center of mass. This center of mass, or *geocenter*, is taken as the origin of a rectangular coordinate system and the rotation axis is used as its z -axis, or x_3 -axis. In this way, neither the earth's figure nor its gravity field nor the coordinate system to which the earth is referred, vary in time.

This simple model is surprisingly accurate, down to an accuracy of 1 part in a million (10^{-6}) and better. The time-variable deformation of the earth because of tides is only of the order of a few decimeters.

Until a decade ago, an accuracy of 10^{-6} was the goal that could realistically be aimed at in the determination of the earth's figure and gravitational field. Then a breakthrough came, mainly from two sides: modern techniques in absolute gravity measurements and laser ranging to satellites achieved accuracies better than 10^{-8} or made them appear feasible. This means accuracies on the order of a few centimeters in absolute position; on this level of precision, geodynamical effects have significant influence.

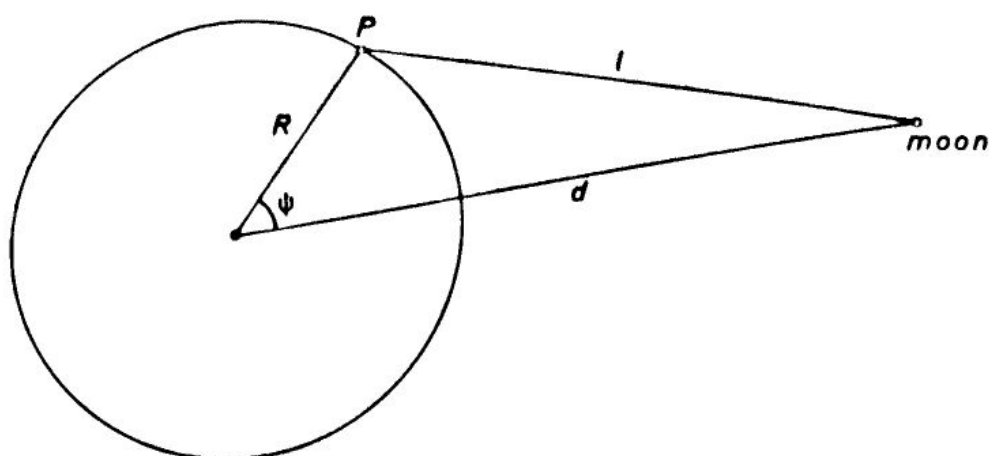
In the present section we can, of course, only sketch the barest outlines. We shall restrict ourselves to two topics: the effect of solid earth tides and some considerations regarding a more precise definition of terrestrial coordinate systems. Other geodynamical effects such as the motion of continental plates according to plate tectonics will not be considered.

We shall primarily discuss geodetic aspects. For geophysical implications, the classical reference is (Munk and Macdonald, 1960); a more recent review article is (Rochester, 1973).

The basic principle is to reduce the observations to the simple rigid earth model mentioned above, on which the usual methods of physical geodesy, as treated in this book, are based. Thus geodynamic effects are taken into account by suitable corrections. In view of the smallness of such corrections, this appears to be the appropriate approach.

Earth tides. Consider the gravitational attraction of the moon at a point P on the earth's surface which, with an accuracy sufficient for the present purpose, can be represented by a sphere of radius R (Fig.55.1). The potential of this attraction at P is

$$v = \frac{G\mu}{r} . \quad (55-1)$$

FIGURE 55.1. *The tidal attraction.*

Expand $1/l$ as a series of spherical harmonics (3-32) to get

$$v = G\mu \sum_{n=0}^{\infty} \frac{R^n}{d^{n+1}} P_n(\cos \psi), \quad (55-2)$$

G being the gravitational constant, μ the mass of the moon; the other notations are understood from Fig. 55.1. The term of zero degree ($n = 0$),

$$v_0 = \frac{G\mu}{d},$$

represents the potential of the attraction of the moon at the center of the earth; it is responsible for orbital motion. The first-degree term ($n = 1$) causes a shift of the equipotential surfaces without changing their shape; only the terms with $n = 2$ and higher correspond to true deformations. Thus the *tidal potential* becomes

$$U = G\mu \sum_{n=2}^{\infty} \frac{R^n}{d^{n+1}} P_n(\cos \psi). \quad (55-3)$$

The dominating term is of second degree. We shall limit ourselves to this term; higher-degree terms can be treated in an analogous fashion if necessary. Thus we shall put

$$U = G\mu \frac{R^2}{d^3} P_2(\cos \psi). \quad (55-4)$$

Let us now express $\cos\psi$ in terms of the geocentric spherical coordinates of P and of the moon's center. In the usual earth-fixed equatorial system the point P has the coordinates (θ, λ) where $\theta = 90^\circ - \phi$ is the polar distance of P , ϕ denoting the geocentric latitude, and λ is the geocentric longitude. Similarly, the moon has the coordinates (p, h) , where the polar distance is given by $p = 90^\circ - \delta$, δ being the declination of the moon, and h denotes the Greenwich hour angle of the moon, that is, the angle between the Greenwich meridian and the meridian passing through the moon's center. Contrary to astronomical usage, both λ and h are counted positively towards east (Fig.55.2).

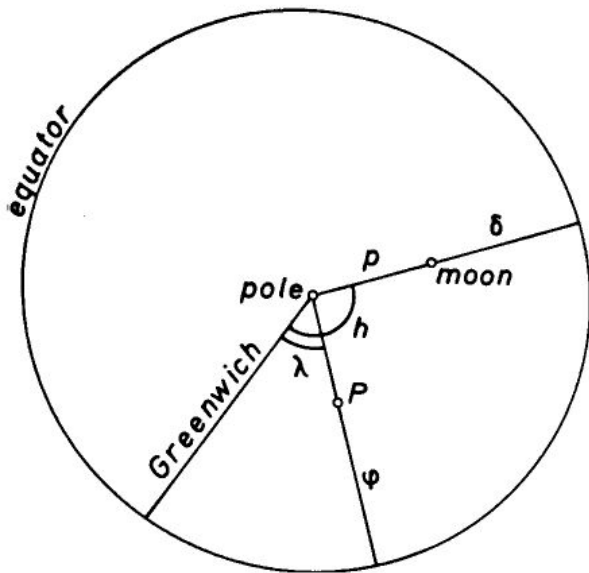


FIGURE 55.2. Coordinates of P and of the moon.

Then the addition theorem of spherical harmonics, eq. (3-30), gives

$$\begin{aligned}
 P_2(\cos\psi) = & R_{20}(\theta, \lambda)R_{20}(p, h) + \\
 & + \frac{1}{3}R_{21}(\theta, \lambda)R_{21}(p, h) + \frac{1}{3}S_{21}(\theta, \lambda)S_{21}(p, h) + \\
 & + \frac{1}{12}R_{22}(\theta, \lambda)R_{22}(p, h) + \frac{1}{12}S_{22}(\theta, \lambda)S_{22}(p, h). \quad (55-5)
 \end{aligned}$$

Now this expression is substituted into (55-4), and all $R_{2m}(p, h)$ and $S_{2m}(p, h)$, as well as the lunar distance d , are represented as functions of time t using the theory of the motion of the moon. Details are found in (Melchior, 1978, chapter 1). The result has the form

$$U = a_1(t)R_{20}(\theta, \lambda) + a_2(t)R_{21}(\theta, \lambda) + a_3(t)S_{21}(\theta, \lambda) + \\ + a_4(t)R_{22}(\theta, \lambda) + a_5(t)S_{22}(\theta, \lambda), \quad (55-6)$$

where the functions $a_i(t)$ can be represented as trigonometric series:

$$a_i(t) = a_{i0} + \sum_{j=1}^{\infty} a_{ij} \cos \omega_j t + \sum_{j=1}^{\infty} b_{ij} \sin \omega_j t. \quad (55-7)$$

A similar expression can be obtained for the effect of the sun, and it will be assumed that (55-6) and (55-7) represent the combined effect of sun and moon.

Thus the tidal potential U at the earth's surface has the form of a linear combination of spherical harmonics of the second degree whose coefficients are quasi-periodic functions of time.

Elastic deformations. If the earth is assumed to be a purely elastic solid, then a point P on its surface undergoes a quasi-periodic displacement on the order of half a meter. This displacement is expressed by the vector \underline{u} whose components in polar coordinates (r, θ, λ) are given by

$$u_r = \frac{h}{g} U, \\ u_\theta = \frac{1}{g} \frac{\partial U}{\partial \theta}, \\ u_\lambda = \frac{1}{g} \frac{\partial U}{\cos \phi \partial \lambda}. \quad (55-8)$$

Here g denotes a mean value of gravity ($g = 980$ gal), and h and l are constants called *Love numbers*.

In the system xyz defined in the usual way (sec.1), the components of the vector \underline{u} are obtained by a rotation:

$$\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} \sin \phi \cos \lambda & -\sin \lambda & \cos \phi \cos \lambda \\ \sin \phi \sin \lambda & \cos \lambda & \cos \phi \sin \lambda \\ -\cos \phi & 0 & \sin \phi \end{bmatrix} \begin{bmatrix} u_\theta \\ u_\lambda \\ u_r \end{bmatrix}. \quad (55-9)$$

By subtracting the tidal deformation \underline{u} from the actual measured coordinates \underline{x}_{obs} , one obtains time-independent coordinates \underline{x} free from tidal effects:

$$\underline{x} = \underline{x}_{\text{obs}} - \underline{u} . \quad (55-10)$$

The deformation of the earth causes its gravitational potential to change by

$$\delta V = kU , \quad (55-11)$$

where k is a third Love number. In space outside the earth, this induced tidal potential is a harmonic function:

$$\begin{aligned} \delta V = k \left(\frac{R}{r} \right)^3 & \left[a_1(t) R_{20}(\theta, \lambda) + \right. \\ & + a_2(t) R_{21}(\theta, \lambda) + a_3(t) S_{21}(\theta, \lambda) + \\ & \left. + a_4(t) R_{22}(\theta, \lambda) + a_5(t) S_{22}(\theta, \lambda) \right] . \end{aligned} \quad (55-12)$$

The fact that both \underline{u} and δV depend linearly on U is due to the smallness of the deformation (any smooth function can for small values of the argument be regarded as a linear function); it is an expression of Hooke's law well known from the theory of elasticity.

Conventional rounded values of the Love numbers are

$$h = 0.6 , \quad k = 0.3 , \quad l = 0.08 . \quad (55-13)$$

Other geodetic effects. The induced potential (55-12) affects satellite orbits (Groten, 1970). Tidal influences on terrestrial observations are usually combined effects of deformation and potential change. Thus, the "geometric" Love numbers h, l usually occur in combination with the "potential" Love number k . For instance, tidal changes in gravity are proportional to the factor $1 + h - 3k/2$; changes in astronomical latitude and longitude involve the factor $1 + k - l$, and horizontal pendulum observations are affected by the factor $1 + k - h$. For details cf. (Melchior, 1971 and 1978) and (Groten, 1979).

Real deformations. The purely elastic model just outlined represents a high degree of idealization. Resonance effects of the liquid outer core of the earth cause a dependence of the Love numbers on frequency ω_j . Such Love numbers can be computed on the basis of an assumed earth model (Melchior, 1978, chapter 6). The picture is further complicated by the fact that there are local disturbances due to the effect of oceanic tides and other local perturbations (*ibid.*, chapters 11 and 12). This makes a model-

ing of tidal effects to an accuracy of a few centimeters quite difficult.

The permanent deformation. The constant term a_{10} in (55-7) is independent of time and causes a "permanent deformation". Actually, only the coefficient a_{10} associated with the zonal harmonic $R_{20}(\theta, \lambda) = P_2(\cos\theta)$ is different from zero, causing a minute change in the flattening of the earth. It has been suggested by Honkasalo (1964) to correct only for the time-dependent part of the tidal effects, leaving the permanent deformation. From a geodetic point of view, it appears, however, preferable to consistently subtract the complete tidal effect including the permanent deformation. In fact, this procedure fully removes the influence of sun and moon and thus provides the simplest way to ensure that the earth's gravitational potential V is harmonic everywhere outside the earth; this is basic for physical geodesy as we have seen throughout the book.

The celestial pole. This concept plays a basic role in the precise definition of terrestrial coordinate systems. We shall be satisfied with a general description of the problem of defining this concept, referring the reader for details to the excellent presentation by A. Leick and I.I. Mueller, "Defining the Celestial Pole", *Manuscripta Geodaetica*, vol.4, pp. 149-183, 1979.

Fig. 55.3 shows the terrestrial sphere in the vicinity of the North Pole as seen from above. The following abbreviations are used:

- F ... figure axis,
- I ... instantaneous rotation axis,
- H ... angular momentum axis,
- E ... Eulerian pole of rotation,
- C ... celestial pole.

The figure axis F (we do not distinguish between an axis and its pole which is its intersection with the celestial or terrestrial sphere) is the axis of maximum inertia (cf. Heiskanen and Moritz, 1967, p.62). If the earth were an ellipsoid of revolution, then F would denote its axis of symmetry.

The axis H corresponds to the direction of angular momentum, which plays a basic role in the dynamics of a rigid body.

For an explanation of the other terms we must distinguish between *free* and *forced motion*. Free motion corresponds to the absence of external forces, in our case, to the absence of gravitational attraction of sun and moon. Forced motion represents the effect of these external forces.

If there were no external forces, then the instantaneous rotation axis I would coincide with E , and the angular momentum axis H would coincide with C . The external forces, however, cause H to describe approximately a circle around C , and I to describe a similar curve around E .

S , such that $OS \approx 2$ m. The points E , C , and S lie on the same radius and slowly rotate together around O ; the period is the Chandler period of polar motion of about 430 days.

So much for the free motion. The attraction of sun and moon causes *forced motions* which may be described as follows. The instantaneous pole of rotation, I , describes a near-circular closed curve around the Eulerian pole E , and the angular momentum pole H performs a similar motion around the celestial pole C , of radii around 0.6 and 0.4 m, respectively. This is of the same magnitude as the tidal deformation: in fact, the cause is the same. The points H and I are again rather close together ($HI \approx 21$ cm).

Especially remarkable is the forced motion of the figure axis F . It describes a quasi-circular motion around its "free" position S ; its radius SF is on the order of 60 m! Thus the figure axis is particularly unstable, its forced motion being a hundred times larger than the forced motion of the rotation axis. It may be mentioned that F , H , and I lie on a straight line.

The period of these forced motions are on the order of 1 day; we therefore speak of *diurnal* polar motions.

The most important point is C , which is therefore called the *celestial reference pole*, or *celestial pole*. By its definition, it is unaffected by forced motion and does not, therefore, exhibit a diurnal motion with respect to an earth-fixed coordinate system. It may also be shown that C has no diurnal motion with respect to a space-fixed system, that is, referred to a nonrotating system which is "fixed with respect to the stars".

These two conditions determine C uniquely and independently of any model assumption such as perfect elasticity; this definition can, therefore, also be used in the case of the real earth. Furthermore it can be shown that most astronomical measurements refer to C rather than to the instantaneous rotation pole I .

This accounts for the use of C to define the celestial pole in the most appropriate way.

The body-fixed motion of C , as considered so far, is the precise definition of *polar motion*; the space-fixed motion of C is *precession and nutation*¹; cf. (Mueller, 1969, sec.4.1).

The new formulas for nutation adopted by the International Astronomical Union at its XVII General Assembly in Montreal in August 1979 refer to the celestial pole C as defined above.

¹"Body-fixed motion" is an abbreviation for "motion with respect to a body-fixed reference system", and similarly for "space-fixed motion".

Terrestrial reference systems. The problem of introducing an appropriate, rigorously defined terrestrial coordinate system is complicated by the fact that there is no such system in which all points on the earth's surface would be at rest. Points are continuously moving because of tidal effects, plate motion, local tectonic disturbances, etc. All that can be hoped for, is that a coordinate system can be defined at which surface points are at rest in some average way.

This problem admits of several solutions and is at present much discussed. The proceedings of a recent meeting on this subject (Kolaczek and Weiffenbach, 1975) give a vivid picture of these discussions.

It is generally agreed that even a future, more precisely defined reference frame should be close to the system presently used: it should be a cartesian coordinate system xyz whose origin is at the geocenter (or perhaps corresponds to some average position of the geocenter); the z -axis should be directed along some average position of the rotation axis, and the zero meridian should be close to the (mean) Greenwich meridian.

Various choices of coordinate axes have distinguished physical properties, as pointed out in (Munk and Macdonald, 1960, pp.10-12).

Tisserand axes. For a rigid body rotating with an angular velocity vector $\underline{\omega}$, the velocity \underline{v} of any particle is

$$\underline{v} = \underline{\omega} \times \underline{x}, \quad (55-14)$$

which is the vector product of $\underline{\omega}$ with the position vector \underline{x} . For a deformable body this relation cannot be satisfied in general; there will be a difference

$$\epsilon = \underline{v} - \underline{\omega} \times \underline{x}. \quad (55-15)$$

If $\underline{\omega}$ is defined by the least-squares condition

$$\iiint_{\text{earth}} \epsilon^2 dM = \text{minimum}, \quad (55-16)$$

dM denoting the element of mass, then any system xyz rotating with this angular velocity is a system of Tisserand axes.

Such a frame also has the property that the total angular momentum due to motion relative to it is zero.

Thus, for Tisserand axes, only the rotation of the frame, that is, its *motion*, is specified. Any cartesian system whose axes are fixed with respect to a Tisserand frame, is also a Tisserand frame.

A Tisserand frame is very suitable for formulating the equations of motion of the earth because they assume a particularly simple form in such a system, being then for a deformable earth formally the same as for a rigid earth. On the other hand, such a system is not directly accessible to geodetic observation.

Principal axes of inertia. These axes are defined in such a way that the inertia tensor is a diagonal matrix in this system: the products of inertia are then zero; cf. (Heiskanen and Moritz, 1967, p.62). They are natural generalizations of axes of symmetry (e.g., for the triaxial ellipsoid) to an arbitrary body and are, therefore, also called *figure axes*.

In the case of the earth, the equatorial principal axes (x and y) are ill defined since the earth is very close to an ellipsoid of revolution.

Even the polar axis of inertia is unsuited for a precise definition of the z -axis because of its instability: the point F to which the polar figure axis corresponds, oscillates with respect to the earth by as much as 60 m, as we have seen above (Fig.55.3).

Mather axes. In his thorough discussion of terrestrial reference frames, Mather (1973, 1974) proposed the following definition. The origin is at the (instantaneous) geocenter; the z -axis coincides with the instantaneous axis of rotation; and one fixed station P on the earth's surface defines the xz -plane (this plane either passes through P , or P has an assigned fixed longitude).

This is perhaps the conceptually clearest and most natural definition of a geodetic reference system. Everything--the origin, the z -axis, and the xz -plane--is unambiguously defined physically. Its main merit lies in presenting a clear theoretical model.

For practical purposes, tidal motions have to be removed by a suitable model, and the z -axis should be oriented through the celestial pole C rather than along the instantaneous axis I because C is observable (see above).

But even so, coordinates xyz of any point on the earth's surface would change with time because of the motion of C with respect to the earth (polar motion) even when there is no real shift of the position of the point under consideration. Furthermore, irregular motions of the fundamental station P would be reflected as spurious temporal changes in the coordinates of the other stations. For this reason it is desirable to define the coordinate system, not with respect to one station P , but with respect to several reference stations, hoping that irregular displacements of the individual stations average out. This leads us to the next definition.

Geographical axes. According to (Munk and Macdonald, 1960, p.11), geographical axes are attached "in a prescribed way" to certain observatories. One possible rigorous definition in this sense would be the following.

Assume N stations (observatories) on the earth's surface. The coordinates $\underline{x}_i = [x_i, y_i, z_i]$, $i = 1, 2, \dots, N$, of these stations, relative to a certain instant t_0 , are given; the coordinate system S_0 is, in principle, arbitrary.

At a subsequent instant t , the rectangular coordinates of the same N stations are again determined by observation, which results in the values $\underline{x}'_i = [x'_i, y'_i, z'_i]$. They are referred to another coordinate system S' which may be arbitrary and unrelated to S_0 .

If the configuration of the N stations did not change with time, then there would be a certain rotation matrix \underline{R} such that

$$\underline{R}\underline{x}' = \underline{x} . \quad (55-17)$$

for all N stations. In view of relative motion of the stations, however, such an equation will not be exactly satisfied; there will be deviations

$$\underline{\epsilon} = \underline{R}\underline{x}' - \underline{x} . \quad (55-18)$$

Now the three parameters defining the rotation matrix \underline{R} (for instance, three Eulerian angles, cf. sec. 36) can be determined by means of the condition

$$\underline{\epsilon}^T \underline{P} \underline{\epsilon} = \text{minimum} \quad (55-19)$$

with a given positive definite weight matrix \underline{P} .¹

This determines the matrix \underline{R} , and now the coordinates \underline{x}' of any point in the new system S' can be transformed to the original system S_0 by (55-17). Thus the coordinates at any instant t can be unambiguously referred to the original system S_0 . This coordinate system is related, not to any physically defined axes, but "in a prescribed way" to the N given observatories.

In the case of errorless observations, the formal least-squares adjustment by means of (55-19) ensures that the configuration of the N stations at epoch t_1 is fitted as closely as possible to the original configura-

¹ The minimum condition (55-19) may, in a way, be considered a discrete analogue of (55-16).

tion of the stations; it is thus a geometrical fitting rather than a statistical adjustment. The "prescribed way" implies the use of the same N stations and of the same matrix \underline{P} at all instants under consideration.

It is clear that tidal effects, etc., which can be represented by an analytical model, should be removed beforehand, so that the residual displacements $\underline{\epsilon}$ in (55-19) have a more or less random character.

In the absence of measuring errors, the residuals (55-18) indicate the amount by which the N given stations have shifted with respect to each other. If there are observational errors, then, of course, the procedure averages their effect as well as the actual displacements. The weight matrix \underline{P} may then be chosen so as to take into account the statistics of the measuring errors.

Practical considerations. Each of these four definitions--Tisserand axes, figure axes, Mather axes, and geographical axes--contains important aspects which must be taken into account in an optimal definition of a terrestrial reference system.

Geographical axes seem best, as far as possible at all, to correspond to the practical requirement that the adopted station coordinates do not change with time. They can also be realized observationally in a theoretically rigorous way.

The arbitrariness of the initial coordinate system S_0 can be used to satisfy the physical requirement that the z -axis is somehow related to the earth's rotation axis. The best candidate for the pole representing the z -axis seems to be a point close to the center O of Fig. 55.3. In the elastic model, the point O represents the long-term average of the rotation axis, of the celestial pole, and of the figure axis; and it does not change its position relative to the earth's body. It can further be shown that O , or any point fixed with respect to it, represents the z -axis of a Tisserand frame.

For the real earth, matters are more complicated than for the elastic model. There is no longer a unique point O which has the same simple physical properties as in the elastic case. Therefore, rather than defining the z -axis *physically* by the point O , it seems appropriate to assume it *conventionally* in such a way that it is close to a mean rotation axis and a mean figure axis. Also the zero meridian will be assumed conventionally but, of course, very close to Greenwich.

The celestial pole can then be referred to this system through polar motion observations by satellite laser, doppler, or VLBI techniques.

The origin of the system S_0 can be placed at the geocenter by dynamical satellite observations or by gravimetric techniques. The basis of the first method is the fact that the geocenter is at the focus of the orbital

ellipse of a satellite (orbital perturbations do not change the principle). The second method uses the phenomenon that a spherical-harmonic expansion of the external potential does not contain first-degree terms if the origin is placed at the geocenter (p.20).

For an *absolute* practical determination of the geocenter with high precision, the satellite technique is appropriate. In this way, the origin of S_0 is placed at the geocenter at the instant t_0 . Since, for subsequent times, this system is defined by reference to N surface stations, it may happen that, at later times t , the origin no longer coincides with the geocenter. For monitoring this *relative* shift of the geocenter, the gravimetric technique may be used, as has been pointed out by Mather (1973, 1974). The principle is as follows.

At any point (ϕ, λ) of the earth's surface, gravity changes because of a shift $\delta \underline{x}$ of the origin by the first-degree harmonic

$$\delta g = c (\cos\phi \cos\lambda \cdot \delta x + \cos\phi \sin\lambda \cdot \delta y + \sin\phi \cdot \delta z) \quad (55-20)$$

where

$$c = -3.1 \text{ } \mu\text{gal cm}^{-1} . \quad (55-21)$$

Monitoring δg by highly precise absolute gravity measurements at a number of well-distributed observatories (preferably coinciding with the N stations used for fitting the coordinate system S_0) thus gives the shift vector $\delta \underline{x}$.

Finally we mention that also the present official terrestrial reference system as defined by the Bureau International de l'Heure (BIH) uses the principle of geographical axes, but fitting astronomical coordinates ϕ and λ of a number of observatories rather than rectangular coordinates (Mueller, 1969, pp.84 and 343).

For a more precise definition, at the centimeter level, satellite laser and interferometric methods using rectangular coordinates are more promising than astronomical coordinates whose accuracy can hardly be essentially improved beyond the present level of a few decimeters. Such future systems should, however, be related to the present system in such a way as to preserve continuity.

REFERENCES

- Backus, G., Inference from inadequate and inaccurate data, I, II, *Proc. Nat. Acad. Sci. U.S.*, 65, 1-7, 281-287, 1970.
- Baeschlin, C.F., *Lehrbuch der Geodäsie*, Orell-Füssli, Zürich, 1948.
- Balmino, G., C. Reigber, and B. Moynot, The Grim 2 earth gravity field model, *Publ. Deut. Geod. Komm.*, A, 86, 1976.
- Berger, M.S., *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.
- Bjerhammar, A., A new theory of geodetic gravity, *Trans. Roy. Inst. Technol.*, 243, Stockholm, 1964.
- Bjerhammar, A., *Theory of Errors and Generalized Matrix Inverses*, Elsevier, Amsterdam, 1973.
- Bjerhammar, A., Discrete approaches to the solution of the boundary value problem in physical geodesy, *Boll. Geod. Sci. Affini*, 34, 185-240, 1975.
- Brosowski, B., and E. Martensen (eds.), *Methoden und Verfahren der mathematischen Physik*, vols. 12, 13, 14 (Mathematical Geodesy), B.I.-Wissenschaftsverlag, Mannheim, 1975.
- Brovar, V.V., On the solution of Molodensky's boundary value problem, *Bull. Geod.*, 72, 167-173, 1964.
- Burkhard, N., and D.D. Jackson, Application of stabilized linear inverse theory to gravity data, *J. Geophys. Res.*, 81, 1513-1518, 1976.
- Clarke, F.L., Three dimensional geodetic network adjustment incorporating gravimetry, *Unisurv G*, 29, 28-81, Univ. of New South Wales, 1978.
- Collatz, L., *The Numerical Treatment of Differential Equations*, Springer, Berlin, 1966.
- Cordova, W.R., New concepts in geodetic instrumentation, in *The Changing World of Geodetic Science* (U.A. Uotila, ed.), Rep. 250, Dep. of Geod. Sci., Ohio State Univ., vol. 1, 73-97, 1977.
- Courant, R., and D. Hilbert, *Methods of Mathematical Physics*, vol. 2, Interscience Publ., 1962.
- Davis, P.J., *Interpolation and Approximation*, Dover Publ., New York, 1975.
- Dermanis, A., Geodetic linear estimation techniques and the norm choice problem, *Manuscripta Geod.*, 2, 15-97, 1977.
- Dermanis, A., Adjustment of geodetic observations in the presence of signals, presented at Int. School of Advanced Geodesy, Erice, Sicily, May-June, 1978 (to appear in *Boll. Geod. Sci. Affini*).
- Deutsch, R., *Estimation Theory*, Prentice-Hall, Englewood Cliffs, N.J., 1965.
- Dieudonné, J., *Foundations of Modern Analysis*, Academic Press, New York, 1960.
- Doob, J.L., Time series and harmonic analysis, in *Proc. Berkeley Symp. Math. Statistics and Probability* (J. Neyman, ed.), 303-343, Univ. of California Press, Berkeley, 1949.
- Ecker, E., The convergence of a series of spherical harmonics, *Boll. Geofis. Teor. Appl.*, 13, 47, 225-233, 1970a.
- Ecker, E., Über die räumliche Konvergenz von Kugelfunktionsreihen, *Publ. Deut. Geod. Komm.*, A, 68, 1970b.
- Ecker, E., Über die Äquivalenz von Lösungen des geodätischen Randwertproblems, *österreich. Z. Vermess.*, 59, 97-105, 1971.

- Ecker, E., *Ausgleichung nach der Methode der kleinsten Quadrate*, Österr. Z. Vermess. Photogramm., 64, 41-53, 1977.
- Eeg, J., and T. Krarup, *Integrated geodesy*, In (Brosowski and Martensen, 1975), vol. 13, 77-123, 1975.
- Erdélyi, A., *Asymptotic Expansions*, Dover Publ., New York, 1956.
- Faddeeva, V.N., *Computational Methods of Linear Algebra*, Dover Publ., New York, 1959.
- Feller, W., *An Introduction to Probability Theory and its Applications*, vol. 1 and 2, Wiley, New York, 1957 and 1966.
- Finetti, B. de, *Theory of Probability*, 2 vols., Wiley, London, 1974.
- Gaposchkin, E.M., *Global gravity field to degree and order 30 from GEOS 3 satellite altimetry and other data*, submitted to J. Geophys. Res., 1979.
- Gentry, D.E., and R.A. Nash, *A statistical algorithm for computing vertical deflections gravimetrically*, J. Geophys. Res., 77, 4912-4919, 1972.
- Giacaglia, G.E.O., and C.A. Lundquist, *Sampling functions for geophysics*, Smithsonian Astrophys. Observatory, Spec. Rep. 344, 1972.
- Gnedenko, B.V., *Theory of Probability*, 4th ed., Chelsea, New York, 1967.
- Gradshteyn, I.S., and I.W. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, New York, 1965.
- Grafarend, E., *A combined gravimetric-astrogeodetic method for telluroid and vertical deflection analysis*, Publ. Deut. Geod. Komm., B, 188, 23-36, 1971.
- Grafarend, E., *Nichtlineare Prädiktion*, Z. Vermess., 97, 245-255, 1972.
- Grafarend, E., *Geodetic prediction concepts*, in (Brosowski and Martensen, 1975), vol. 13, 161-200, 1975.
- Grafarend, E., *Geodetic applications of stochastic processes*, Phys. Earth Planet. Interiors, 12, 151-179, 1976.
- Grafarend, E., *Space-time differential geodesy*, in *The Changing World of Geodetic Science* (U.A. Uotila, ed.), Rep. 250, Dep. of Geod. Sci., Ohio State Univ., vol. 1, 150-216, 1977.
- Grafarend, E., *Operational geodesy*, in (Moritz and Sünkel, 1978), 235-284, 1978.
- Grafarend, E., and G. Offermanns, *Eine Lotabweichungskarte Westdeutschlands nach einem geodätisch konsistenten Kolmogorov-Wiener-Modell*, Publ. Deut. Geod. Komm., A, 82, 1975.
- Groten, E., *On tidal effects in satellite gravity data*, Comm. Obs. Royal Belg., A, 9, 228-233, 1970.
- Groten, E., *Geodesy and the Earth's Gravity Field*, 2 vols., Dümmler, Bonn, 1979.
- Groten, E., G. Hein, and H. Jochemczyk, *On the determination of empirical covariances*, Allg. Vermess. Nachr., 17-32, 1979.
- Hardy, R.L., *Geodetic applications of multiquadric equations*, Rep., Engineering Res. Inst., Iowa State Univ., 1976.
- Heiskanen, W.A., and H. Moritz, *Physical Geodesy*, W.H. Freeman, San Francisco, 1967. (Reprint 1979 by Institute of Physical Geodesy, Technical University, Graz, Austria; also available from Department of Geodetic Science, Ohio State University, Columbus.)
- Heitz, S., and C. Tscherning, *Comparison of two methods of astrogeodetic geoid determination based on least-squares prediction and collocation*, Tellus, 24, 271-276, 1972.
- Helmert, F.R., *Die mathematischen und physikalischen Theorien der höheren Geodäsie*, vol. 2, B.G. Teubner, Leipzig, 1884 (reprinted 1962).
- Hobson, E.W., *The Theory of Spherical and Ellipsoidal Harmonics*, Cambridge Univ. Press, 1931.
- Honkasalo, T., *On the tidal gravity correction*, Boll. Geofis. Teor. Appl., VI, 34-36, 1964.

- Hörmander, L., *An Introduction to Complex Analysis in Several Variables*, Van Nostrand, Princeton, N.J., 1966.
- Hörmander, L., The boundary problems of physical geodesy, *Arch. Rat. Mech. Anal.*, 62, 1-52, 1976.
- Hotine, M., *Mathematical Geodesy*, ESSA Monograph 2, U.S. Dep. of Commerce, Washington, 1969.
- IAG, Geodetic Reference System 1967, Spec. Publ., *Bull. Geod.*, 1970.
- Jekeli, C., An investigation of two models for the degree variances of global covariance functions, Rep. 275, Dep. of Geod. Sci., Ohio State Univ., 1978.
- Jordan, S., Statistical model for gravity, topography, and density contrasts in the earth, *J. Geophys. Res.*, 83, 1816-1824, 1978.
- Kantorovich, L.V., and G.P. Akilov, *Functional Analysis in Normed Spaces*, Macmillan, New York, 1964.
- Kaula, W.M., Statistical and harmonic analysis of gravity, *J. Geophys. Res.*, 64, 2401-2421, 1959.
- Kaula, W.M., Determination of the earth's gravitational field, *Rev. Geophys. Space Phys.*, 1, 507-551, 1963.
- Kaula, W.M., Global harmonic and statistical analysis of gravimetry, in *Gravity Anomalies: Unsurveyed Areas*, *Geophys. Monogr. Ser.*, vol. 9, 58-67, Am. Geophys. Union, Washington, 1966.
- Kaula, W.M., Theory of statistical analysis of data distributed over a sphere, *Rev. Geophys. Space Phys.*, 5, 83-107, 1967.
- Kearsley, W., Non-stationary estimation in gravity prediction problems, Rep. 256, Dep. of Geod. Sci., Ohio State Univ., 1977.
- Kellogg, O.D., *Foundations of Potential Theory*, Springer, Berlin, 1929 (reprinted 1967).
- Knopp, K., *Theorie und Anwendung der unendlichen Reihen*, 5th ed., Springer, Berlin, 1964.
- Koch, K.R., Solution of the geodetic boundary-value problem in case of a reference ellipsoid, Rep. 104, Dep. of Geod. Sci., Ohio State Univ., 1968.
- Koch, K.R., Die geodätische Randwertaufgabe bei bekannter Erdoberfläche, *Z. Vermess.*, 96, 218-224, 1971.
- Koch, K.R., and A.J. Pope, Uniqueness and existence for the geodetic boundary value problem using the known surface of the earth, *Bull. Geod.*, 106, 467-476, 1972.
- Kolaczek, B., and G. Weiffenbach (eds.), *Proc. IAU Colloquium on Reference Coordinate Systems for Earth Dynamics*, Toruń (Poland), August, 1974.
- Kolmogorov, A.N., and S.V. Fomin, *Introductory Real Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1970.
- Kostelecký, J., and J. Klokočník, 30th order harmonics from resonant inclination variations of ten satellites, *Observations of Artificial Satellites of the Earth*, 18, 227-231, *Pol. Acad. Sci.*, Warsaw, 1978.
- Krarpup, T., A contribution to the mathematical foundation of physical geodesy, Publ. 44, *Dan. Geod. Inst.*, Copenhagen, 1969.
- Krarpup, T., Letters on Molodensky's Problem I-IV, Communication to the members of IAG Special Study Group 4.31, unpublished, 1973.
- Krarpup, T., On potential theory, in (Brosowski and Martensen, 1975), vol. 12, 79-160, 1975.
- Krarpup, T., Some remarks about collocation, in (Moritz and Sünkel, 1978), 193-209, 1978.
- Kreyszig, E., *Introductory Mathematical Statistics*, Wiley, New York, 1970.
- Kryński, J., Improvement of the geoid in local areas by satellite-to-satellite tracking, *Bull. Geod.*, 53, 19-36, 1979.

- Lachapelle, G., Determination of the geoid using heterogeneous data, Publ. Geod. Inst. Tech. Univ. Graz, 19, 1975.
- Lachapelle, G., A spherical harmonic expansion of the isostatic reduction potential, Boll. Geod. Sci. Affini, 35, 281-299, 1976.
- Lachapelle, G., Estimation of disturbing potential components using a combined integral formulae and collocation approach, Manuscripta Geod., 2, 233-262, 1977.
- Landkof, N.S., Foundations of Modern Potential Theory, Springer, Berlin, 1972.
- Lauritzen, S., The probabilistic background of some statistical methods in physical geodesy, Publ. 48, Dan. Geod. Inst., Copenhagen, 1973.
- Lavrentiev, M.M., Some Improperly Posed Problems of Mathematical Physics, Springer, Berlin, 1967.
- Ledersteger, K., Astronomische und physikalische Geodäsie (Erdmessung), vol. V of Jordan/Eggert/Kneissl, Handbuch der Vermessungskunde, J.B. Metzler, Stuttgart, 1969.
- Leigemann, D., Untersuchungen zu einer genaueren Lösung des Problems von Stokes, Publ. Deut. Geod. Komm., C, 155, 1970.
- Leigemann, D., Spherical approximation and the combination of gravimetric and satellite data, Boll. Geod. Sci. Affini, 32, 241-250, 1973.
- Leigemann, D., Zur gravimetrischen Berechnung des Geoides der Bundesrepublik Deutschland, Publ. Deut. Geod. Komm., A, 77, 1974.
- Leigemann, D., Ein Verfahren zur astro-gravimetrischen Geoidbestimmung, Publ. Deut. Geod. Komm., C, 247, 1978.
- Lerch, F.J., S.M. Klosko, R.E. Laubscher, and C.A. Wagner, Gravity model improvement using GEOS 3 (GEM 9 & 10), Publ. X-921-77-246, Goddard Space Flight Center, Greenbelt, Md., 1977.
- Levallois, J.J., Géodésie Générale, vol. 3, Eyrolles, Paris, 1970.
- Levallois, J.J., Remarques générales sur la convergence du développement du potentiel terrestre en harmoniques sphériques, Boll. Geod. Sci. Affini, 32, 53-77, 1973.
- Liebelt, P.B., An Introduction to Optimal Estimation, Addison-Wesley, Reading, Mass., 1967.
- Loomis, L.H., and S. Sternberg, Advanced Calculus, Addison-Wesley, Reading, Mass., 1968.
- Magnizki, W.A., W.W. Browar, and B.P. Schimbirew, Theorie der Figur der Erde, VEB Verlag für Bauwesen, Berlin, 1964.
- Marych, M.I., On the second approximation of M.S. Molodensky for the disturbing potential (in Russian), Geodezija, Kartografija i Aerofotosyemka, 10, 17-27, Lvov Univ., 1969.
- Mather, R.S., Four dimensional studies in earth space, Bull. Geod., 108, 187-209, 1973.
- Mather, R.S., Time variations in geodetic coordinates, Unisurv G, 20, 77-132, Univ. of New South Wales, 1974.
- Meissl, P., A study of covariance functions related to the earth's disturbing potential, Rep. 151, Dep. of Geod. Sci., Ohio State Univ., 1971a.
- Meissl, P., On the linearization of the geodetic boundary value problem, Rep. 152, Dep. of Geod. Sci., Ohio State Univ., 1971b.
- Meissl, P., Preparations for the numerical evaluation of second order Molodensky-type formulas, Rep. 163, Dep. of Geod. Sci., Ohio State Univ., 1971c.
- Meissl, P., Elements of functional analysis, in (Brosowski and Martensen, 1975), vol. 12, 19-78, 1975.
- Meissl, P., Hilbert spaces and their application to geodetic least-squares problems, Boll. Geod. Sci. Affini, 35, 49-80, 1976.
- Melchior, P., Physique et Dynamique planétaires, 4 vols., Vander, Louvain, 1971.

- Melchior, P., *The Tides of the Planet Earth*, Pergamon Press, Oxford, 1978.
- Meschkowski, H., *Hilbertsche Räume mit Kernfunktion*, Springer, Berlin, 1962.
- Mikhail, E.M., *Observations and Least Squares*, Dun-Donnelley, New York, 1976.
- Mikhlin, S.G., *Multidimensional Singular Integrals and Integral Equations*, Pergamon Press, Oxford, 1965.
- Miranda, C., *Partial Differential Equations of Elliptic Type*, 2nd ed., Springer, Berlin, 1970.
- Mises, R. von, *Probability, Statistics, and Truth*, 2nd ed., Allen and Unwin, London, 1957.
- Molodenskii, M.S., V.F. Eremeev, and M.I. Yurkina, *Methods for Study of the External Gravitational Field and Figure of the Earth*, Transl. from Russian (1960), Israel Program for Scientific Translations, Jerusalem, 1962.
- Monti, C., and F. Sansò, Applicazione del metodo della collocazione all'analisi dell'errore di graduazione del cerchio di un teodolite, *Boll. Geod. Sci. Affini*, 36, 215-234, 1977.
- Moritz, H., Über die Konvergenz der Kugelfunktionsentwicklung für das Außenraumpotential an der Erdoberfläche, *Österr. Z. Vermess.*, 49, 11-15, 1961.
- Moritz, H., Zur geometrischen Deutung der Minimumsprinzipien der Ausgleichungsrechnung, *Z. Vermess.*, 91, 293-296, 1966.
- Moritz, H., Linear solutions of the geodetic boundary-value problem, *Publ. Deut. Geod. Komm.*, A, 58, 1968a.
- Moritz, H., On the use of the terrain correction in solving Molodensky's problem, Rep. 108, *Dep. of Geod. Sci., Ohio State Univ.*, 1968b.
- Moritz, H., A general theory of gravity processing, Rep. 122, *Dep. of Geod. Sci., Ohio State Univ.*, 1969a.
- Moritz, H., Nonlinear solutions of the geodetic boundary-value problem, Rep. 126, *Dep. of Geod. Sci., Ohio State Univ.*, 1969b.
- Moritz, H., Least-squares estimation in physical geodesy, *Publ. Deut. Geod. Komm.*, A, 69, 1970.
- Moritz, H., Series solutions of Molodensky's problem, *Publ. Deut. Geod. Komm.*, A, 70, 1971.
- Moritz, H., Convergence of Molodensky's series, Rep. 183, *Dep. of Geod. Sci., Ohio State Univ.*, 1972.
- Moritz, H., Least-squares collocation, *Publ. Deut. Geod. Komm.*, A, 75, 1973a.
- Moritz, H., On the convergence of Molodensky's series, *Boll. Geod. Sci. Affini*, 32, 125-144, 1973b.
- Moritz, H., Precise gravimetric geodesy, Rep. 219, *Dep. of Geod. Sci., Ohio State Univ.*, 1974.
- Moritz, H., Integral formulas and collocation, *Manuscripta Geod.*, 1, 1-40, 1976a.
- Moritz, H., Covariance functions in least-squares collocation, Rep. 240, *Dep. of Geod. Sci., Ohio State Univ.*, 1976b.
- Moritz, H., On the computation of a global covariance model, Rep. 255, *Dep. of Geod. Sci., Ohio State Univ.*, 1977a.
- Moritz, H., Recent developments in the geodetic boundary-value problem, Rep. 266, *Dep. of Geod. Sci., Ohio State Univ.*, 1977b.
- Moritz, H., Least-squares collocation, *Rev. Geophys. Space Phys.*, 16, 421-430, 1978a.
- Moritz, H., Introduction to interpolation and approximation, in (Moritz and Sünkel, 1978), 1-45, 1978b.

- Moritz, H., Statistical foundations of collocation, Rep. 272, Dep. of Geod. Sci., Ohio State Univ., 1978c.
- Moritz, H., The operational approach to physical geodesy, Rep. 277, Dep. of Geod. Sci., Ohio State Univ., 1978d.
- Moritz, H., On the convergence of the spherical-harmonic expansion for the geopotential at the earth's surface, *Boll. Geod. Sci. Affini*, 37, 363-381, 1978e.
- Moritz, H., and K.P. Schwarz, On the computation of spherical harmonics from satellite observations, *Boll. Geod. Sci. Affini*, 32, 185-200, 1973.
- Moritz, H., and H. Sünkel (eds.), *Approximation Methods in Geodesy*, H. Wichmann, Karlsruhe, 1978.
- Morrison, F., Azimuth-dependent statistics for interpolating geodetic data, *Bull. Geod.*, 51, 105-118, 1977.
- Mueller, I.I., *Spherical and Practical Astronomy*, F. Ungar, New York, 1969.
- Munk, W.H., and G.J.F. Macdonald, *The Rotation of the Earth*, Cambridge Univ. Press, 1960.
- Nash, R.A., and S.K. Jordan, Statistical geodesy - an engineering perspective, *Proc. IEEE*, 66, 532-550, 1978.
- Nashed, Z., Approximate regularized solutions to improperly posed linear integral and operator equations, in *Constructive and Computational Methods for Differential and Integral Equations* (D.L. Colton and R.P. Gilbert, eds.), *Lecture Notes in Mathematics*, vol. 430, Springer, Berlin, 1974.
- Natanson, L.P., *Theorie der Funktionen einer reellen Veränderlichen*, Akademie-Verlag, Berlin, 1969.
- Neyman, Yu.M., Probabilistic modification of Stokes' formula for the computation of height anomalies (in Russian), *Geod. Aerofotosyemka*, 21-24, 1974.
- Neyman, Yu.M., On the regularization of Molodensky's boundary value problem (in Russian), *Geod. Aerofotosyemka*, 57-64, 1975.
- Neyman, Yu.M., A variational method of solving discrete problems of physical geodesy (in Russian), *Geod. Aerofotosyemka*, 21-27, 1977.
- Papoulis, A., *Systems and Transforms with Applications in Optics*, McGraw-Hill, New York, 1968.
- Parzen, E., An approach to time series analysis, *Ann. Math. Statist.*, 32, 951-989, 1961.
- Pellinen, L.P., Accounting for topography in the calculation of quasigeoidal heights and plumb-line deflections from gravity anomalies, *Bull. Geod.*, 63, 57-65, 1962.
- Pellinen, L.P., On the identity of various solutions of Molodensky's problem with the help of a small parameter, *Int. Symp. on Earth Gravity Models and Related Problems*, Saint Louis, Missouri, August, 1972 (also *Geod. Aerofotosyemka*, 65-71, 1974).
- Pellinen, L.P., *Higher Geodesy* (in Russian), Nedra, Moscow, 1978.
- Pick, M., J. Picha, and V. Vyskočil, *Theory of the Earth's Gravity Field*, Elsevier, Amsterdam, 1973.
- Poincaré, H., *Les méthodes nouvelles de la Mécanique Céleste*, vol. 2, Gauthier-Villars, Paris, 1893.
- Rapp, R.H., Numerical results from the combination of gravimetric and satellite data using the principles of least-squares collocation, Rep. 200, Dep. of Geod. Sci., Ohio State Univ., 1973.
- Rapp, R.H., Comparison of least-squares and collocation estimated potential coefficients, in (Brosowski and Martensen, 1975), vol. 14, 133-148, 1975.
- Rapp, R.H., Potential coefficient determinations from 5° terrestrial gravity data, Rep. 251, Dep. of Geod. Sci., Ohio State Univ., 1977.

- Rapp, R.H., Results of the application of least-squares collocation to selected geodetic problems, in (Moritz and Sünkel, 1978), 117-156, 1978.
- Rochester, M.G., The earth's rotation, EOS Trans. Am. Geophys. Union, 54, 769-780, 1973.
- Rummel, R., A model comparison in least-squares collocation, Bull. Geod., 50, 181-192, 1976.
- Sagrebín, D.W., Die Theorie des regularisierten Geoids, Publ. Geod. Inst. Potsdam, 9, 1956.
- Sansó, F., Discussion on the existence and uniqueness of the solution of Molodensky's problem in gravity space, Accad. Naz. Lincei, Rend. Sci. fis. mat. e nat., 61, 260-268, 1976.
- Sansó, F., The geodetic boundary value problem in gravity space, Mem. Accad. Naz. Lincei, 14, 39-97, 1977.
- Sansó, F., Molodensky's problem in gravity space: a review of the first results, Bull. Geod., 52, 59-70, 1978a.
- Sansó, F., The local solvability of Molodensky's problem in gravity space, Manuscripta Geod., 3, 157-227, 1978b.
- Sansó, F., The minimum mean square estimation error principle in physical geodesy, presented at 7th Symp. on Mathematical Geodesy, Assisi, June, 1978c (to appear in Boll. Geod. Sci. Affini).
- Schwartz, J.T., Nonlinear Functional Analysis. Gordon and Breach, New York, 1969.
- Schwarz, K.P., Application of collocation: Spherical harmonics from satellite observations, in (Brosowski and Martensen, 1975), vol. 14, 111-132, 1975.
- Schwarz, K.P., Least-squares collocation for large systems, Boll. Geod. Sci. Affini, 35, 309-324, 1976a.
- Schwarz, K.P., Geodetic accuracies obtainable from measurements of first and second order gravitational gradients, Rep. 242, Dep. of Geod. Sci., Ohio State Univ., 1976b.
- Schwarz, K.P., Capabilities of airborne gradiometry for gravity estimation, Boll. Geod. Sci. Affini, 36, 195-214, 1977.
- Schwarz, K.P., On the application of least-squares collocation models to physical geodesy, in (Moritz and Sünkel, 1978), 89-116, 1978a.
- Schwarz, K.P., Geodetic improperly posed problems and their regularization, Lecture Notes, International School of Advanced Geodesy, Erice (Sicily), May-June, 1978b (to appear in Boll. Geod. Sci. Affini).
- Schwarz, K.P., and J. Kryński, Improvement of the geoid in local areas by satellite gradiometry, Bull. Geod., 51, 163-176, 1977.
- Shimbirov, B.P., Theory of the Figure of the Earth (in Russian), Nedra, Moscow, 1975.
- Sjöberg, L., A comparison of Bjerhammar's methods and collocation in physical geodesy, Rep. 273, Dep. of Geod. Sci., Ohio State Univ., 1978.
- Smirnow, W.I., Lehrgang der Höheren Mathematik, vol. III/1, 4th ed., VEB Deutscher Verlag der Wissenschaften, Berlin, 1964a.
- Smirnow, W.I., Lehrgang der Höheren Mathematik, vol. III/2, 5th ed., VEB Deutscher Verlag der Wissenschaften, Berlin, 1964b.
- Smirnow, W.I., Lehrgang der Höheren Mathematik, vol. II, 7th ed., VEB Deutscher Verlag der Wissenschaften, Berlin, 1966.
- Smirnow, W.I., Lehrgang der Höheren Mathematik, vol. I, 8th ed., VEB Deutscher Verlag der Wissenschaften, Berlin, 1967.
- Sternberg, S., Celestial Mechanics, vol. 2, W.A. Benjamin, New York, 1969.
- Sünkel, H., Die Darstellung geodätischer Integralformeln durch bikubische Spline-Funktionen, Publ. Geod. Inst. Tech. Univ. Graz, 28, 1977.

- Sünkel, H., Zur Geometrie des normalen Schwerfeldes, *Österr. Z. Vermess. Photogramm.*, 66, 71-85, 1978a.
- Sünkel, H., Approximation of covariance functions by non-positive definite functions, Rep. 271, Dep. of Geod. Sci., Ohio State Univ., 1978b.
- Taylor, A.E., *Introduction to Functional Analysis*, Wiley, New York, 1958.
- Tienstra, J.M., *Theory of the Adjustment of Normally Distributed Observations*, Argus, Amsterdam, 1956.
- Tikhonov, A.N., and V.Y. Arsenin, *Solutions of Ill-Posed Problems*, Wiley, New York, 1977.
- Torge, W., *Geodäsie*, de Gruyter, Berlin, 1975.
- Tscherning, C.C., Representation of covariance functions related to the anomalous potential of the earth using reproducing kernels, Internal Rep. 3, Dan. Geod. Inst., Copenhagen, 1972.
- Tscherning, C.C., A Fortran IV program for the determination of the anomalous potential using stepwise least-squares collocation, Rep. 212, Dep. of Geod. Sci., Ohio State Univ., 1974.
- Tscherning, C.C., Application of collocation for the planning of gravity surveys, *Bull. Geod.*, 116, 183-198, 1975a.
- Tscherning, C.C., Application of collocation: Determination of a local approximation to the anomalous potential of the earth using "exact" astro-gravimetric collocation, in (Brosowski and Martensen, 1975), vol. 14, 83-110, 1975b.
- Tscherning, C.C., Covariance expressions for second and lower order derivatives of the anomalous potential, Rep. 225, Dep. of Geod. Sci., Ohio State Univ., 1976.
- Tscherning, C.C., A note on the choice of norm when using collocation for the computation of approximations to the anomalous potential. *Bull. Geod.*, 51, 137-147, 1977.
- Tscherning, C.C., Introduction to functional analysis with a view to its applications in approximation theory, in (Moritz and Sünkel, 1978), 157-191, 1978a.
- Tscherning, C.C., Collocation and least squares methods as a tool for handling gravity field dependent data obtained through space research techniques, *Bull. Geod.*, 52, 199-212, 1978b.
- Tscherning, C.C., and R.H. Rapp, Closed covariance expressions for gravity anomalies, geoid undulations, and deflections of the vertical implied by anomaly degree variance models, Rep. 208, Dep. of Geod. Sci., Ohio State Univ., 1974.
- Wiener, N., Generalized harmonic analysis, *Acta Math.*, 55, 117-258, 1930.
- Wiener, N., *Time Series*, M.I.T. Press, Cambridge, Mass., 1949.
- Wladimirow, W.S., *Gleichungen der mathematischen Physik*, VEB Deutscher Verlag der Wissenschaften, Berlin, 1972.
- Wolf, H., *Ausgleichsrechnung nach der Methode der kleinsten Quadrate*, Dümmler, Bonn, 1968.
- Wolf, H., Die Sonderfälle der diskreten Kollokation, *Österr. Z. Vermess. Photogramm.*, 65, 132-138, 1977.
- Wolf, H., *Ausgleichsrechnung II*, Dümmler, Bonn, 1979.
- Zygmund, A., *Trigonometric Series*, vol. 1, Cambridge Univ. Press, 1968.

INDEX

- Addition theorem, 23
- Adjustment, least-squares, 111, 117, 144, 167, 220, 251
- Analytical continuation, 54, 96, 377
- Analytical continuation solution, 377, 400, 419
- Angular momentum, 482
- Atmospheric effects, 422

- Banach space, 44, 434
- Bjerhammar problem, 95
- Bjerhammar sphere, 69, 181
- Boundary condition, fundamental, 15, 342, 352, 443
- Boundary-value problem
 - fixed, 437, 450, 475
 - free, 437, 450, 475
 - geodetic, 330
- Brillouin sphere, 430
- Brovar solution, 365, 396, 401
- Bruns formula, 14, 342, 354, 424

- Cartesian product, 222
- Cauchy sequence, 41
- Celestial pole, 484
- Centrifugal force, 4
- Codimension, 207
- Collocation, 85, 257
 - analytical, 93, 220
 - least-squares, 85, 99, 111, 132, 207, 249, 313
 - sequential, 155
 - statistics of, 255, 297, 307
 - stepwise, 144, 150
- Completeness, 41
- Convergence, absolute, 406
- Convergence, sphere of, 57
- Coordinates
 - astronomical, 5
 - geodetic, 7
 - natural, 7
 - rectangular, 2, 485
 - spherical, 18
 - terrestrial, 2, 485
- Correlation length, 174
- Covariance function, 81, 169, 181, 263, 283
 - empirical, 94, 189, 266, 283
- Covariance matrix, 76, 161
- Covariance propagation, 86, 105

- Data combination, 143, 221
- Deflection of the vertical, 8, 106, 419
- Delta function, 29, 258
- Dense, 42
- Dirichlet problem, 333, 402, 430, 441, 452, 473
- Distributions, statistical, 297
- Dynamic form factor, 12

- Earth tides, 477
- Ecker formula, 394
- Ellipsoid, 7, 10, 52, 337
- Ellipsoidal corrections, 314, 425
- Ergodicity, 269, 285
- Equipotential surface, 6
- Error covariance matrix, 77, 125, 142, 154,
- Estimate, linear, 77, 122
- Euler angles, 288
- Euler equation, 244
- Evaluation functional, 37, 200

- Faye anomaly, 419
- Figure axes, 486
- Filtering, 105, 133
- Fourier series, 34, 260
- Fréchet derivative, 435
- Fredholm alternative, 429, 432
- Full rank, 77
- Functional, 36, 50, 222
 - evaluation, 37, 200
 - linear, 36, 45

- Geocenter, 2, 477, 489
- Geodynamical effects, 477
- Geographical axes, 487
- Geoid, 6
- Geoidal height, 9
- Geopotential number, 7
- Gradient, 4
- Gradient solution, 387, 414, 421
- Gradient tensor, 4
- Gradient variance, 176
- Gravitational constant, 3
- Gravitational field, 3
- Gravity, 5
 - normal, 11
- Gravity anomaly, 14, 338, 353
- Gravity change, 489
- Gravity field, 3

Gravity reduction, 311
 Gravity space, 449
 Gravity vector, 4
 Green identity, 394

Hahn-Banach theorem, 70
 Hankel transformation, 178
 Harmonic function, 4
 Height, orthometric, 6
 Height anomaly, 353, 364, 425
 Hilbert space, 25, 44, 196
 Hölder norm, 412, 438
 Hooke law, 481
 Hörmander theorems, 433, 447

Improperly posed problem, 239, 313
 Integral equation, singular, 403
 Inverse, generalized, 164, 240
 Inverse problem, geophysical, 169, 313
 Inverse function problem, 434, 475
 Isozenithal, 345

Keldysh-Lavrentiev theorem, 65, 69
 Kelvin transformation, 58, 463
 Kernel of an integral operator, 26
 Kernel function, 93, 196
 Kolmogorov-Wiener formula, 80
 Krarup linearization, 337
 Kronecker delta, 27

Laplace equation, 4, 461
 Laplacian, 4, 461
 Lauritzen theorem, 270, 285, 308
 Legendre function, 19
 Legendre polynomial, 19
 Legendre transformation, 452
 Level surface, 6
 Lipschitz condition, 412
 Love number, 480

Mapping, 47
 Marussi condition, 450, 452
 Marussi telluroid, 338, 353
 Mather axes, 486
 Measurements, geodetic, 135, 221
 Model approach, 221
 Molodensky problem, 330, 349, 351, 440, 449, 451
 existence and uniqueness, 428, 447, 474
 simple, 351, 430
 Molodensky series, 364, 401
 Molodensky shrinking, 360, 370, 380, 404

Nash-Hörmander iteration, 435
 Newton method, 335, 437, 475
 Noise, 99
 Norm, 24, 30, 38
 Nutation, 484

Oblique-derivative problem, 349, 403, 457, 472
 Observation equations, 134, 230
 Occam's razor, 309
 Operational approach, 221
 Orthogonality relations, 21, 260
 Orthonormal functions, 23, 31
 Operator, 25, 49
 bounded, 38
 inverse, 27
 linear, 25, 38
 unit, 27

Pellinen identity, 390
 Permanent tidal deformation, 482
 Planar approximation, 170, 359
 Plumb line, 5
 Poisson equation, 4
 Polar motion, 484
 Positive definite, 79, 177, 196, 206
 Potential
 adjoint, 452
 anomalous (disturbing), 12, 230, 33
 gravitational, 2
 gravity, 3
 induced tidal, 481
 normal, 10, 230, 337
 tidal, 478
 Potential anomaly, 338
 Precession, 484
 Prediction, 80, 105, 133
 gravity, 80
 least-squares, 80
 Principal axes, 486
 Product, inner, 30, 44, 215

Random variable, 263
 Reference systems, 2, 331, 485, 488
 Representer of a functional, 198
 Reproducing kernel, 196, 201
 Riesz representation theorem, 46
 Rotation axis, 2, 482
 Rotation group, 288, 302, 313
 Runge theorem, 64, 67, 98, 388
 Sansō problem, 450
 Satellite observations, 138, 156, 225
 Sectorial harmonics, 20
 Sensitivity matrix, 111

500 *Index*

- Signal, 99
- Singular integral, 402
- Somigliana formula, 11
- Space
 - dual, 198
 - gravity, 449
 - inner product, 44
 - linear, 43
 - normed, 43
 - product, 222
- Spectrum, 178, 197
- Spherical approximation, 15, 82, 106, 316, 351, 462
- Spherical harmonics, 20
 - convergence of, 50, 63
 - determination of, 156
 - fully normalized, 22
- Stochastic process, 261, 279, 308
- Stokes formula, 15, 98, 364, 379, 419, 425
- Stokes function, 17, 366
- Stokes problem, 15, 330, 428
- Telluroid, 335, 337, 353
 - gravimetric, 339, 457
- Tensor calculus, 203
- Terrain correction, 414
- Tesseral harmonics, 20
- Tides, 477
- Tikhonov regularization, 240, 248
- Tisserand axes, 485
- Topographic surface, 6
- Variational principle, 243
- Vening Meinesz formula, 17, 379, 425
- Vertical, 5
- Very-long-baseline interferometry, 228
- Zonal harmonics, 20